partial equilibrium setting, and assume that Pareto conditions are always satisfied elsewhere in the economy. We also ignore income effects and assume risk neutrality so that expected consumer's surplus can serve as an indicator of welfare. The objective is to choose prices P_i^* for $i = 1, 2, \ldots, n$, and capacity z^*, to maximize a welfare function equal to the expected consumer's surplus plus revenue (i.e., willingness to pay) less expected cost (i.e., variable cost plus capacity cost), subject to the constraint that total revenue will just equal total cost. In treating the welfare-maximizing problem an assumption will be needed about the willingness to pay of those who are actually served whenever quantity demanded exceeds capacity. Such a question was of no interest to the expected profit-maximizing monopolist because profit, unlike consumer's surplus, is not affected by the assumption.

In this section we assume that the service produced is always distributed efficiently in the sense that consumers who value it most are the ones who receive the service, whether or not money price is high enough to ensure that result. This "efficient nonprice rationing" assumption was introduced by Brown and Johnson and also relied upon by Meyer. The assumption is interesting. When demand is stochastic and there is no break-even constraint, it leads to welfare-maximizing prices that equal short-run marginal costs (Brown-Johnson). But these prices do not yield enough revenue to cover total cost (although there are no economies of scale in production) and so a balanced budget-constrained second best welfare analysis is relevant. In Sections III and IV we explore the implications of relaxing this efficient nonprice rationing assumption.

There is an important contrast to be anticipated as we move from certainty to the stochastic demand case. At the optimum solution under certainty, excess demand would never occur. Indeed, with the cost function assumed here, the budget constraint always could be satisfied under certainty with efficient market-clearing prices (as, for example, in Williamson), and the budget constraint therefore would be inessential, or redundant. When demand is random, though, we have no assurance that any price or capacity level chosen before demand is known will clear the market once demand is revealed. And as we just noted, Brown and Johnson have shown that if efficient nonprice rationing is assumed in this situation, the welfare-maximizing set of prices and capacity will not allow the producer to break even. For if prices are set high enough to break even on average, capacity will go unutilized whenever demand is unusually low. Some consumers who valued the service above its marginal cost would be denied service then, and the outcome would be inefficient. To avoid such a result first best prices are set at short-run marginal cost, lower than break-even prices. But that means a break-even constraint will be binding in the stochastic demand problem even though it was not binding in the certainty model.

Let us now add a stochastic element to the demand function and derive a solution to the balanced budget constrained, second best problem. Assume that, as in the expected profit-maximizing model, demand in period i is given by the function $X_i(P_i) + u_i$ where u_i is distributed with density function $f_i(u_i)$ and mean zero. Again, period i lasts a fraction α_i of the demand cycle, and z^* and the P_i^*'s are unchanging values to be chosen and held at the same levels regardless what actual values of u_i occur. Total expected consumer's surplus and revenue is

$$\sum_{i=1}^{n} \alpha_i \left\{ \int_{x}^{\infty} f_i(u_i) \cdot \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i] dP_i du_i + P_i^* X_i(P_i^*) \right\}$$

less the expected loss in both consumer's surplus and revenue because some consumers are not served (in this case those who value service least) when quantity demanded exceeds capacity:

$$\sum_{i=1}^{n} \alpha_i \left\{ \int_{z^* - X_i(P_i^*)}^{\propto} f_i(u_i) \cdot \right.$$

$$\int_{P_i^*}^{X_i^{-1}(z^*-u_i)} [X_i(P_i) + u_i - z^*]dP_i du_i +$$

$$P_i^* \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,[X_i(P_i^*) + u_i - z^*]du_i \Bigg\}$$

Variable costs are

$$\sum_{i=1}^{n} \alpha_i b \left\{ X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \cdot [X_i(P_i^*) + u_i - z^*]du_i \right\}$$

and capacity cost is βz. The balanced budget constraint requires that

$$\sum_{i=1}^{n} \alpha_i (P_i^* - b)\left\{ X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \cdot [X_i(P_i^*) + u_i - z^*]du_i \right\} = \beta z^*$$

We can thus construct a second best problem, maximizing expected consumer's surplus plus revenue less variable and capacity costs subject to the break-even constraint, by forming the Lagrangian

$$(8)\quad L(P_i^*, z^*, \lambda) = \sum_{i=1}^{n} \alpha_i \Bigg\{ \int_{-\infty}^{\infty} f_i(u_i) \cdot \int_{P_i^*}^{X_i^{-1}(\cdot u_i)} [X_i(P_i) + u_i]\,dP_i du_i$$

$$+ P_i^* X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \int_{P_i^*}^{X_i^{-1}(z^*-u_i)} \cdot [X_i(P_i) + u_i - z^*]\,dP_i du_i - P_i^* \int_{z^*-X_i(P_i^*)}^{\infty}$$

$$f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i - b[X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i \; - \beta z^* \Bigg\}$$

$$+ \lambda \left\{ \sum_{i=1}^{n} \alpha_i (P_i^* - b)[X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i] - \beta z^* \right\}$$

Differentiating the Lagrangian with respect to each P_i^* and z^*, and setting results equal to zero, we have

$$(9)\quad \frac{\partial L}{\partial P_i^*} = \alpha_i \Bigg\{ \lambda X_i(P_i^*) - \lambda \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i(P_i^*) + u_i - z^*]\,du_i + (1 + \lambda)$$

$$P_i^* - b) X_i'(P_i^*) F(z^* - X_i(P_i^*)) \Bigg\} = 0$$

$$(10)\quad \frac{\partial L}{\partial z^*} = \sum_{i=1}^{n} \alpha_i \Bigg\{ \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i^{-1}(z^* - u_i) - P_i^*]\,du_i + (P^* - b) \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,du_i \Bigg\} - \beta$$

$$+ \lambda \left\{ \sum_{i=1}^{n} \alpha_i (P_i^* - b) \cdot \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,du_i - \beta \right\} = 0$$

Remember from (4) above our definition of demand elasticities η_i, and from (6) the definition of expected excess demand $E(ed_i)$. We can obtain from (9) optimum pricing rules in the form

$$(11)\quad \frac{(P_i^* - b)F(z^* - X_i(P_i^*))}{P_i^*} = \frac{\lambda}{1 + \lambda} \cdot \frac{1}{\eta_i}\left[1 - \frac{E(ed_i)}{X_i(P_i^*)}\right] \qquad i = 1, \ldots, n$$

Condition (11) is a constrained expected welfare-maximizing counterpart to the expected profit-maximizing implicit pricing rule in equation (5). Comparison of equation (11) with equation (5) shows that the only difference in form between the second best expected welfare-maximizing profit margin $(P_i^* - b)/P_i^*$ and the expected profit-maximizing profit margin is the constant term $\lambda/(1 + \lambda) < 1$ at the right-hand side of (11).

The relation between the pricing rules for monopoly (5) and for welfare (11) goals involves only a constant on the right-hand side as in the relation without risk illustrated by Baumol and Bradford. This is not surprising when one realizes that short-run marginal cost b is the first best price here if

THE AMERICAN ECONOMIC REVIEW

VOLUME LXVIII

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: Brown University, Providence, Rhode Island

CONTENTS OF ARTICLES AND SHORTER PAPERS

CONTENTS OF PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

THE AMERICAN ECONOMIC REVIEW

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this *REVIEW* is assumed by the editors or the publishers, The American Economic Association.

March 1978

VOLUME 68, NUMBER 1

Articles

Shorter Papers

Number 79 of a series of photographs of past presidents of the Association

Lawrence R. Klein

The Supply Side

By LAWRENCE R. KLEIN*

I. The Meaning of Supply and Demand in a Macroeconomic Context

It is worth considering whether a new basic model should guide our thinking about performance of the economy as a whole. It is not that the macro models of the past twenty-five years or so have failed to serve us well. When we consider the state of our knowledge about the analytics of the economy at the end of World War II and the apprehensiveness with which we approached the modern era of expansion, it should be evident that we have come a long way professionally. Yet the economic problems of today seem to be intractable when studied through the medium of simplified macro models. The new system should combine the Keynesian model of final demand and income determination with the Leontief model of interindustrial flows. This is the motivation for my focusing attention on the supply side of the economy.

It is frequently said, in almost an offhand manner, that the theories of aggregate employment and output determination are demand models, that economic policy for overall direction of the economy is a policy of *demand* management. I would generally agree with these remarks, but not in every last detail, once the meaning of demand in these contexts is carefully pulled apart and analyzed. The *demand* aspects are possibly overstated.

It is, of course, true that demand for the *GNP* built up as the sum of demands by consumers, businesses, government, and foreigners (consumption, investment, public spending, and net exports) covers total demand in the economy and is composed of demands by the constituent parts. But demand by firms, and, in many cases by government, are not ends in themselves. Business demand is largely for goods to produce goods. The capital formation that results from business demand goes into the increment of capital stock, after allowance for capital consumption, and the capital stock becomes a factor input in the production function. The accumulation of capital contributes to the *supply* of goods and services. Indeed, investment *demand* now for new capital facilitates the implementation of the production process with the supply of factors of ever-increasing powers of productivity, thus making it possible to supply increasing amounts of goods and services with inputs that are increasing at a somewhat slower rate.

By focusing attention excessively on the "short run," in which the capital stock is timelessly held fixed by assumption only and not in reality, we have ignored the supply-side characteristics of investment demand. Students of today's business cycle commonly cite investment demand as the promising potential route to higher productivity in the relatively near future, thereby lessening inflationary pressure. In this respect, economic theoreticians have been myopic relative to the applied economic analysts in the world of affairs. Nevertheless, as we shall see, there is much more to the supply side than the transformation of investment into productive capital, and the basic characterization of contemporary macroeconomics as demand analysis has a point. A strong indication of the demand side orientation is given by the elaboration of the standard macro model. In place of aggregate consumption, the more elaborate model gives separate treatment to consumer expenditures on durables, nondurables, and services. This is the first stage. At a higher stage, there is further disaggregation into types of durables, nondurables, and services such as food, cars, medical services, etc. The detail that is introduced for consump-

*Presidential address delivered at the ninetieth meeting of the American Economic Association, New York City, New York, December 29, 1977.

tion is repeated in business investment demand, housing demand, public expenditures, exports, and imports. Elaboration essentially means taking a closer look at demand side components by types of demand.

The *mainstream* model of macro economic thought has thus become a detailed system of demand analysis, but if it is to be a closed system, it will also have to include corresponding detail on the national income side of the social accounts. If this is done fully, there will have to be analyses of factor rewards, factor use, and pricing. The development of factor demand goes beyond capital formation, which appears as a demand for final goods in the *GNP*, and takes up an explanation of wage income. An adequate explanation of wage income cannot avoid the explicit treatment of physical production involving labor input as well as capital input. The demand for labor, like the demand for capital, is supply side analysis. While the demand for capital enters directly as a component of total demand, the demand for labor, together with wage formation, enters national income, and only after expenditure does it enter final demand for *GNP*. To the extent that labor productivity affects wage determination and also price formation, we find supply-side factors influencing inflation and consequently the overall performance of the economy. Labor demand can also be associated with training. The training component is, in fact, investment—in human rather than fixed capital. Looked at in this way, factor demand for labor and factor demand for fixed capital are simply different, but related, aspects of total investment.

Behind the *IS-LM* diagram or other simplified renditions of the aggregate demand model lie many supply-side relationships. Not only is the supply side in the background, but it also plays a more essential role once it is recognized that the simplified model is actually incomplete. If we were to assume the existence of money illusion, it would be possible to consider the *IS-LM* system as a closed system of relationships depending on nominal income and nominal interest rates. I find this approach theoretically unsatisfactory. That simple system exists only as an aggregative approximation for a given price level. If we assume no money illusion and, more properly, I believe, the need to determine the aggregate price level, then the *IS-LM* diagram does not provide a closed system analysis, and we must extend the system to include the whole supply-side apparatus of production relationships, factor demand, and factor supply.

It is well known that Keynes included the aggregate supply function in the General Theory, but it was introduced in his chapter on "The Principle of Effective Demand." That part of his analysis dealing with supply has been largely played down by the profession at large—not by all students of macroeconomics.[1] Also, by way of side comment, Keynes probably confused the issues by making labor supply dependent on the nominal wage rate, assuming the existence of money illusion, and by not treating the stock of capital as an explicit variable.

If the demand relationships explaining the components of the *GNP* are disaggregated into a highly detailed set, it does not necessarily mean that the supply side must be equally disaggregated to a similar extent, as long as the total flow of income and purchasing power to be directed towards the expenditure flow can be generated. The detailed expenditure flow will, however, involve price relatives. That is a consequence of disaggregation. An aggregate supply-side explanation that generates only an average price level for output as a whole can be adequate, provided separate prices, needed for the price-relatives, can be explained in terms of a relationship to the overall price or wage level. This is much like the use of a term structure relationship in credit market analysis to explain the spectrum of interest rates, given one strategic rate.

It is, however, more satisfactory, and more revealing, to explain the whole set of prices, one by one, on the basis of costs in individual sectors. These sector prices, on the side of production, are then combined with input weights into the several final de-

[1] See Sidney Weintraub (1956, 1957).

mand prices needed to account for variation in components of final expenditure. This brings us to a fundamental set of new considerations on the supply side.

II. The Task of Modeling Supply

If sector prices by line of production are to be explained in a fundamental way by sector costs, there will have to be an accompanying explanation of sector outputs and inputs. This brings us directly to the supply side of things. While the supply side is represented in the macro production function from an aggregative point of view, once we disaggregate the supply side by sector of production, we encounter a new dimension. The aggregate production function, in the spirit of Paul Douglas and Charles Cobb, expresses *value-added* as a function of primary factor inputs, namely, labor and capital. They were able to compress the technology as they did, because at a full macro level, one sector's output is someone else's input, and for the economy as a whole, only value-added is left in the output aggregation. Intermediate inputs or outputs may be neglected in the interests of avoiding double counting. This way of looking at things is strictly correct only for a closed economy. In an open system, intermediate imports must be treated like primary factor inputs.

At the sector level, however, there is no question about the need to consider intermediate inputs. Sector output (gross) is properly a function of intermediate inputs, labor input, and capital input—all sector designated. The presently fashionable way of summarizing this idea is to use the *KLEM* production function, whose inputs consist of capital, labor, energy, and materials.

The *KLEM* production function concept is useful in partial studies of separate industries or sectors, and has long been anticipated in aggregative production function studies. It has been routine in production function studies in agriculture to use feed, seed, fertilizer, and other intermediate inputs as explanatory variables. The dependent variable is generally a measure of gross output—gallons of milk, bushels of grain, or bales of cotton. In manufacturing, one of the earliest studies was by Ragnar Frisch. He expressed isoquants for the output of the Freia chocolate factory as a function of fat content and molding-cooling input. One of these is a pure material input and the other stands for some capital, labor, and general running cost input. In my own investigations of *U.S.* railroad production functions, I included fuel consumption (in coal equivalents) as one of the factor inputs together with labor and capital. The gross output concept consisted of a log-linear combination of ton miles and passenger miles.[2]

These individual industry production functions with a small number of intermediate inputs are hardly substitutes for a detailed input-output analysis on a general system level. The role of input-output analysis is to explain *intermediate* flows in the economic system. The full system is needed in order to provide an adequate supply analysis because

(i) There is much more to economic activity than can be summarized by the system of final goods production.

(ii) The explanation of types of final prices depends on highly specific types of intermediate, as well as final, goods/services prices.

The occurrence of bottlenecks—potential or realized—as in the oil embargo of 1973–74 or the diversion of large amounts of agricultural output to export markets as in 1973 and 1975 are striking examples of cases where there was a great deal of economic activity going on outside final *GNP* sectors. An economic understanding of those activities and an estimate of their macro impacts on the *GNP* could not be readily derived from demand analysis without consulting the table of intermediate flows in *I-O* analysis. These are only striking examples. Many more have arisen in the past, and more are bound to occur in the future; therefore, the concern of this presentation is not with singular events.

An adequate explanation of the price system, especially on the cost side, cannot stop

[2]See the author.

at the *KLEM* level with separate consideration of energy, materials, wage, and capital costs. It must take account of prices of grains, ferrous metals, nonferrous metals, coal, crude oil, machinery, textiles, and the other component prices in an input-output system. The appropriate amount of detail is not a fixed matter. It depends on human capabilities of analysis, machine facilities, data bases, and other practical considerations, but it is, in any case, an order of magnitude greater than contemplated by mainstream macro model analysis.

From an analytical point of view, what is being suggested is a full combination of two systems of thought, the Leontief model and the Keynesian model. That these two systems can be put back-to-back into a single consistent model, with full feedback between each part, is now well known, having been implemented first with the Brookings Model and later with various generations of Wharton Models, and more recently by Dale Jorgenson in a translog mode. A principal feature of such combined systems is that they are not based on restrictive assumptions of the fixed coefficient input-output model, but are generalized to allow the coefficients of production to vary, according to the variation of relative prices.

The above expression, "full feedback," means that the macro model of final demand and national income generation cannot be solved, by itself, without also solving the input-output system for generating sector production flows. Moreover sector prices cannot be solved without also solving the macro model simultaneously.

Price formation in individual sectors is specified in terms of mark-up relations over unit labor costs. Thus, sector outputs and labor inputs are needed in order to explain sector prices. These prices are needed, in turn, in order to explain final demand prices. Similarly, sector investment depends on sector output as well as sector price. It is for these and similar reasons that final demand cannot be generated without making use of the input-output system in order to generate sector outputs.

At the same time, the input-output system is driven by final demand; therefore, the conventional macro demand model must be used in order to solve the input-output system. These are the specific senses in which full feedback is used in order to obtain simultaneous and consistent integration of the entire supply and demand sides of the economy.

In terms of the history of economic thought, the above approach means thinking in terms of the empirical implementation of the Walrasian system. Essentially, Tinbergen implemented the Keynesian system and Leontief implemented a part of the Walrasian system. By putting the two together, with due allowance to Kuznets for making the data bases of final demand and national income available, a complete synthesis of supply and demand in the economy as a whole can be put together. This gives the antecedents of what is meant by modeling supply, taking into account what is needed from demand models at the same time.

III. Why Model Supply?

At the time of the Keynesian Revolution, there was a pervasive deficiency of demand throughout most of the world. The Keynesian policy development, building on that model, did, in my opinion, much good for the economy of the Western world, enabling us to come through an expansive era of more than twenty-five years without a recurrence of a Great Depression. That does not mean that this system of thought and policy formation did its work for all time in putting the world economy on a stable footing. It carried the situation only so far, and undoubtedly underestimated inflation potentials, leaving us now at the point where new systems of thought, drawing more on the supply side, are needed in order to develop policies that will be able to deal with the world's contemporary economic problems; hopefully, policies that will have as much longevity as the demand management policies of the last two to three decades. That should bring us nicely into the twenty-first century, which is about as far ahead as we might attempt to look at the present time.

The limits of demand management policies have become clearly visible in recent years. Let us look at the issues through the medium of specific problems, say the joint problems of too much unemployment and too much inflation. Policies of demand management alone have appeared to be adequate to deal with one or the other, but not both together. If demand is stimulated enough to bring down the unemployment rate to a full-employment minimum, there is danger of generating undue inflationary pressure as a side effect. Conversely, anti-inflationary policies of demand restriction run the danger of generating excessive unemployment while holding down the inflation rate.

How might supply-side policies be introduced to lower both the inflation and unemployment rates at the same time? It is conventionally thought that policies of aggregative demand stimulus through traditional fiscal and monetary policies might be able to bring down the *U.S.* unemployment rate to about 5.5 percent. This is not a firm point estimate, and is subject to error of at least one-half point above or below that figure, but it is not, in any case, a full-employment target figure.

One way, but not the only way, of getting to full employment without generating fresh inflationary pressure is to design a jobs program for about 1.0 million long-term, hard core unemployed. This jobs program cannot be described in full detail in the context of this presentation, but it is not to be viewed as an ordinary public jobs program. It is viewed as a job training program aimed at people who show signs of receptivity to training and enlisting the participation of employers who provide really productive jobs with potential for upward mobility. The 1.0 million target, spread over three years, is not purely indicative. It is meant to be plausible and necessary if full employment is to be reached by 1980 82 in the United States.

Apart from the fact that some public funds are to be spent on this program, it is not a typical demand management policy. It is aimed at increasing the supply of goods, at raising labor productivity, at sectors of the economy where job training can be accomodated or needed, and at sectors of the labor force. It is basically a supply-side policy and needs for its implementation/assessment a full scale analysis through the medium of a Leontief-Keynes system. In first approximations, such assessments have been made with the appropriate version of the Wharton Model.

In anticipation of criticism of this policy approach from the side of those who are strongly wedded to emphasis on demand management, I want to stress that a jobs program aimed at increasing productivity and reducing hard core unemployment is not a futile exercise in pushing some subsidized workers into the ranks of the employed while pushing others out. The program is intended to have balance; i.e., to be part of a larger program with corresponding support from the demand side. Such support could not be justified from the point of view of inflation potential unless steps are being taken to complement the effect with a jobs program and eventual lifting of productivity. Undue preoccupation with demand policies is not going to be adequate to meet the problems of the day, nor is pure emphasis on supply. Both sides of the economy must be coordinated in policy formation.

It should also be emphasized that demand policies of federal expenditures for public service employment appear to be inferior to private sector jobs programs of the type being mentioned here. In the former case, there is no long-term opportunity for those taken into the program and there is no contribution to national productivity. As long as job expenditures are going to be made, they ought, preferably, to be directed to an effort that promises to have some lasting benefit.

This example of the jobs program is one that fits the contemporary American economic scene and has been investigated with a *U.S.* model and data. The underlying idea, however, is meant to be much more general. It is that the whole industrial world is faced with a series of new supply-side economic problems. The problems of cyclical stabilization and reaching full em-

ployment without inflation will have to be dealt with as before, and the latter will require some degree of supply-side analysis in other economies as in the *U.S.* case, but a whole new range of economic issues looms on the horizon. These are development of new, greater energy supplies, protection of the environment, controlling the exhaustion of resources, enhancing agricultural supplies, balancing population development, and others of like nature. The juggling of public budgets, the setting of tax rates, and the giving of a tone to money market conditions are not going to deal effectively with this new class of problems, from the viewpoints of influencing them in a favorable direction. Similarly, the demand oriented model is not going to provide much understanding of them.

The coming problems of the industrial economy are not going to be wholly dealt with or analyzed on the basis of the general purpose Leontief-Keynes system that is being advocated here. In many cases, the unforeseen problems that are bound to arise are going to be more specialized than can be conveniently anticipated. In such cases, the analysis must extend into partial system analysis giving more detailed and explicit treatment on the supply side. In terms of model building that means construction of many "satellite" systems on the supply side, as the need arises. At the present time, many energy satellite systems are being developed to deal with new fuel processes, large energy using sectors, and large energy delivery sectors. These satellite systems are then all linked, in a technical and consistent way, with the input-output cum macro model system. In any event, the intent is to move the discussion of macroeconomics and policy formation to a new plane of discourse.

The discussion, thus far, has focused on the modeling and related policy problems for the modern industrial economy. The analysis of the supply side, however, is not a new issue for the developing economy. A deficiency of demand analyzed within the framework of the Keynesian Model has not generally been thought to be the issue or approach for dealing with the problems of economic development. That is not to say that demand relations are nonexistent or unimportant for the developing economy. It is primarily a matter of emphasis. Availability of fixed capital treated as a limiting factor in production is central to understanding the problems facing many developing economies.

Energy problems of production and use are already apparent, as are population control and agricultural production. Where problems of environmental protection and resource exhaustion have not yet arisen, they are bound to occur in significant instances; therefore, it is wise for the development economist to be forearmed with a full model for analysis of both supply and demand sides.

The centrally planned economies are for the most part industrial economies and have the same needs for supply-side analysis. In their cases, the supply side has perhaps been excessively developed with inadequate attention paid to the demand side, not from the viewpoint of deficient demand but from the viewpoint of chronic excess demand, with latent inflationary pressure.

The present analysis attempts to look at a particular facet of the modern economy, namely the supply side. That does not imply, by any means, that monetary analysis and policy are unimportant. Most of the supply-side problems have monetary implications or aspects; therefore, monetary policy must be appropriate to insure the smooth working of the supply side of the economy.

In terms of the analytical apparatus needed to combine monetary analysis with the kind of supply-demand model that I have outlined above, it is a matter of integrating the flow-of-funds system together with the input-output and final demand national income system. It would also be in a full feedback mode. To complement the supply-side detail underlying the *IS* curve, we would turn to the complete flow-of-funds model to provide background for the *LM* curve.

It is my feeling that overall monetary and fiscal policies have been overworked, with expectations of results that are not justified.

Without downgrading their very important role, the present message simply says that a full supply-side analysis must be developed into which an elaborated *IS-LM* system of thought can be fully integrated.

REFERENCES

R. Frisch, "The Principle of Substitution: An Example of its Application in the Chocolate Industry," *Nordisk Tidsskrift for Teknisk Økonomi*, Sept. 1935, *1*, 21-27.

K. C. Hoffman and D. W. Jorgenson, "Economic and Technological Models for Evaluation of Energy Policy," *Bell J. Econ.*, Autumn 1977, *8*, 444-66.

Lawrence R. Klein, *A Textbook of Econometrics*, Evanston 1953.

S. Weintraub, "A Macroeconomic Approach to the Theory of Wages," *Amer. Econ. Rev.*, Dec. 1956, *46*, 835-56.

———, "The Micro-Foundations of Aggregate Demand and Supply," *Econ. J.*, Sept. 1957, *67*, 455-70.

The Economics of "Tagging" as Applied to the Optimal Income Tax, Welfare Programs, and Manpower Planning

By GEORGE A. AKERLOF*

The advantages of a negative income tax are easy to describe. Such a tax typically gives positive work incentives to even the poorest persons. With some forms of the negative income tax there are no incentives for families to split apart to obtain greater welfare payments. Furthermore, individuals of similar income are treated in similar fashion, and therefore it is fair and also relatively cheap and easy to administer.

In contrast to these advantages of a negative income tax, the advantages of a system of welfare made up of a patchwork of different awards to help various needy groups are less easy to describe and also less well understood. Such a system uses various characteristics, such as age, employment status, female head of household, to identify (in my terminology to "tag") groups of persons who are on the average needy. These groups are then given special treatment, or, as the economist would view it, they are given a special tax schedule different from the rest of the populace. A system of tagging permits relatively high welfare payments with relatively low marginal rates of taxation, a proposition which will be explained presently and discussed at some length.

I

It is the aim of this paper to explore the nature of the optimal negative income tax with tagging and to compare this tax with the optimal negative income tax in which all groups are treated alike. I should emphasize at the outset, however, that I do not wish to defend one type of welfare system versus another—rather, I feel that if welfare reform is to be successful, the merits of different systems must be understood, especially the merits of the system which is to be replaced. The evidence is fairly strong that the proponents of welfare reform have failed to understand (or to face) the costs involved in going from a system of welfare based on tagging (such as we now have in the United States) to one which treats all people uniformly.

The role of tagging in income redistribution can be seen most simply in a very simple formula and its modification. Consider a negative income tax of the form $T = -\alpha\bar{Y} + tY$, where α is the fraction of per capita income received by a person with zero gross income, t is the marginal rate of taxation, and $\bar{Y}$ is per capita income. Summing the left-hand side and the right-hand side of this formula over all individuals in the economy and dividing by total income yields a formula of the form:

$$t = \alpha + g \tag{1}$$

where g is the ratio of net taxes collected to total income, and t and α come from the formula for the negative income tax.[1] For-

*Professor, University of California-Berkeley. I am indebted to George Borts and an anonymous referee for invaluable comments. I would also like to thank the National Science Foundation for research support under grant number SOC 75-23076, administered by the Institute of Business and Economic Research, University of California-Berkeley.

[1] Define g as: $\Sigma T_i / \Sigma Y_i$, where g is net tax collections relative to total income. Formula (1) can be derived as follows: $T_i = -\alpha\bar{Y} + tY_i$ is the taxes paid by individual i. Summing over all i individuals (assumed to be n in number),

$$\sum_{i=1}^{n} T_i = \sum_{i=1}^{n} -\alpha\bar{Y} + \sum_{i=1}^{n} tY_i$$

$$\sum_{i=1}^{n} T_i = -\alpha n\bar{Y} + t\sum_{i=1}^{n} Y_i \tag{a}$$

Because $\bar{Y}$ is by definition, $(\Sigma Y_i)/n$, and because g is

mula (1) indicates the fundamental tradeoff involved in income redistribution by a linear negative income tax. Higher levels of support α can be given, but only at the cost of higher marginal rates of taxation. Thus, if α is 40 percent and g is 15 percent, numbers which are not unrealistic, marginal tax rates are 55 percent.

Suppose, however, that it is possible to identify (tag) a group which contains all the poor people and that this group contains only a fraction β of the total population. By giving this tagged group a minimum support, which is a fraction α of average income and a marginal tax rate t, and by giving untagged persons a zero support level and the same marginal tax rate t, similar to formula (1), we find:[2]

$$(2) \qquad t = \beta\alpha + g$$

Formula (2) shows that tagging makes the tradeoff between levels of support and marginal rates of taxation more favorable by eliminating the grant to taxpayers, and thus allows greater support for the poor with less distortion to the tax structure.

Table 1 is taken from the 1974 *Economic Report of the President* (p. 168). This table indicates the scope and magnitude, and also the importance, of tagging in federal redistribution programs. Such programs as aid to the aged, the blind, and the disabled, and also Medicare (including such aid administered by the Social Security system), are examples of tagging. Such programs as aid to families with dependent children are less clearcut—but it must be remembered that this program began as Aid to Dependent Children, and assistance was given to families with children without able-bodied fathers.

Female-headed households have a particularly high incidence of poverty, and this criterion (despite its perverse incentive to families to split up) was therefore one of the most efficient techniques of tagging. Other programs, such as Medicaid and housing subsidies, represent a form of tagging most common in underdeveloped and Communist countries. Since poor people spend a greater fraction of their income on some items than others, the subsidization of items of inferior but utilitarian quality constitutes one method of income "redistribution." It is also an example of tagging. In sum, Table 1 shows, to a fairly good degree of accuracy, that *U.S.* federal redistribution schemes are, with some exceptions, based on tagging.

Furthermore, the record of the debate on welfare reform reveals that the central issues involve the tradeoffs between α, t, and β reflected in formulas (1) and (2). Recall that, in August 1969, President Nixon proposed the Family Assistance Plan. By this

---

by definition, $\sum_{i=1}^{n} T_i / \sum_{i=1}^{n} Y_i$, a division of the left-hand and the right-hand sides of (a) by ΣY_i yields:

$$\frac{\Sigma T_i}{\Sigma Y_i} = -\alpha \frac{n\bar{Y}}{\Sigma Y_i} + t$$

whence, $g = -\alpha + t$, and $t = \alpha + g$.

[2]Formula (2) is derived in similar fashion to formula (1). Let n_p denote the number of poor people, with $n_p/n = \beta$. (Let poor people be numbered 1 to n_p.) Poor people pay a tax

$$T_i = (-\alpha\bar{Y} + tY_i) \qquad i = 1, \ldots, n_p$$

whereas other people pay a tax

$$T_i = tY_i \qquad i = n_p + 1, \ldots, n$$

Thus, total net revenues are:

$$\sum_{i=1}^{n} T_i = \sum_{i=1}^{n_p} (-\alpha\bar{Y} + tY_i) + \sum_{i=n_p+1}^{n} tY_i$$

and

$$\sum_{i=1}^{n} T_i = -n_p\alpha\bar{Y} + t\sum_{i=1}^{n} Y_i$$

or using the definition of β, $n_p = \beta n$

$$(b) \qquad \sum_{i=1}^{n} T_i = -\beta\alpha n\bar{Y} + t\sum_{i=1}^{n} Y_i$$

Dividing the left-hand and right-hand sides of (b) by ΣY_i yields:

$$\frac{\sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} Y_i} = -\beta\alpha \frac{n\bar{Y}}{\sum_{i=1}^{n} Y_i} + t$$

or $g = -\beta\alpha + t$.

TABLE 1—FEDERAL GOVERNMENT TRANSFER PROGRAMS, FISCAL YEAR 1973

Program	Total Expenditure (millions of dollars)	Number of Recipients (thousands)	Monthly Benefits per Recipient[a]	Percent of Recipients in Poverty[b]
Social Security				
Old age and survivors insurance	42,170	25,205	$139	16
Disability insurance	5,162	3,272	132	24
Public Assistance				
Aid of families with dependent children	3,617	10,980	c	76
Blind	56	78	c	62
Disabled	766	1,164	c	73
Aged	1,051	1,917	c	60
Other Cash Programs				
Veterans' compensation and benefits	1,401	7,203	74	(4)
Unemployment insurance benefits	4,404	5,409	68	(4)
In Kind				
Medicare	9,039	10,600	71	17
Medicaid	4,402	23,537	c	70
Food stamps	2,136	12,639	14	92
Public housing	1,408	3,319	c	d
Rent supplements	106	373	24	d
Homeownership assistance (section 235)	282	1,647	14	d
Rental housing assistance (section 236)	170	513	28	d

[a]The number of recipients is for individuals, not families.
[b]Poverty is defined relative to money income and the size of the recipient's family. Money income includes money transfer payments but excludes income received in kind. All percents are estimated.
[c]Programs with federal-state sharing of expenses.
[d]Not available.

plan a typical welfare family would receive $1,600 per year if it earned no income at all (*New York Times*, Aug. 9, 1969). There would be no decrease in benefits for the first $720 earned, but thereafter a 50¢ decline in benefits for every dollar earned up to an income of $3,920. The debate on this proposal in Congress was long and discussed many peripheral questions, but one central issue stands out. On the one side were those, with Senator Abraham Ribicoff as the leading protagonist, who considered the benefits too "meager" (Ribicoff's phrase, *New York Times*, Apr. 21, 1970); on the other side was the administration, with a succession of secretaries of Health, Education, and Welfare as leading protagonists, who viewed any increase in these benefits as too "costly" (Elliott Richardson's phrase, *New York Times*, July 22, 1971). By this it was meant that with such an increase the marginal tax rate t would have to be too great. No compromise was reached, and in March 1972 the bill was withdrawn by the administration. In the background, of course, was the current welfare system, whose tagging programs allow a better tradeoff between α and t—even though other incentives such as incentives to work and to maintain a family may be perverse.

Thus, formula (1) and its modification with tagging are instructive and pertain to real issues. These formulas are generally useful in showing the two-way tradeoff between welfare support and marginal rates of taxation, and the three-way tradeoff between these two variables and tagging. It is fairly intuitive by consumer's surplus arguments that the cost of a tax is the "deadweight loss" due to the gap created between private and social marginal products, which in this case is the marginal rate of taxation itself; ideally, however, the welfare cost of a tax is endogenous and should be derived from basic principles of utility maximization and general equilibrium analysis.

Ray Fair and James Mirrlees have developed the theory of the negative income tax uniformly applied. Their approach is reviewed in the next section, because, with added complication, the tradeoffs may be applied to a model of the optimal negative income tax with tagging. Section III illustrates the proposition that tagging of poor people typically results in greater support levels to the poor. Section IV gives a complicated and generalized model of optimal income redistribution with tagging, of which Section III presented a simple but illustrative example. Section V discusses the relation between tagging and the estimation of costs and benefits of manpower programs. Section VI gives conclusions.

II. A Simple Example and Explanation of Mirrlees-Fair

Following the example of Mirrlees and Fair, there is a population with a distribution of abilities a, according to the distribution function $f(a)$. Members of this population receive income dependent on their marginal products of the form $w(a)L(a)$, where $w(a)$ is the wage of a worker of ability of index a, and $L(a)$ is the labor input of such a worker. After-tax income is $w(a)L(a) - t(w(a)L(a))$, where $t(y)$ is the tax paid on gross income y. Members of this population have utility positively dependent on after-tax income and negatively dependent on labor input. Thus, utility of a person of ability a is

$$u(a) = u[w(a)L(a) - t(w(a)L(a)), L(a)] \tag{3}$$

The optimal tax is defined as maximizing the expected value of the utility of the population, denoted U,

$$U = \int u[w(a)L(a) - t(w(a)L(a)), L(a)]f(a)da \tag{4}$$

subject to the constraint that taxes equal transfers, or,

$$\int t(w(a)L(a))f(a)da = 0 \tag{5}$$

and also subject to the constraint that each individual chooses his labor input to maximize his utility, given the wage rate paid to persons of his ability, his utility function u, and the tax schedule $t(y)$, yielding the first-order condition:

$$\frac{\partial}{\partial L(a)}\{u[w(a)L(a) - t(w(a)L(a)), L(a)]\} = 0$$

However complicated the equations or the mathematics, the basic tradeoff made in the choice of an optimal Mirrlees-Fair style income tax can be explained as follows. As taxes are raised and incomes are redistributed, there is a gain in welfare, because income is distributed to those who have greater need of it (higher marginal utility). But this gain must be balanced against a loss: as tax rates rise in relatively productive jobs and as subsidies rise in relatively unproductive jobs, workers are less willing to take the productive (and more willing to take the unproductive) jobs. Such switching, per se, results in a loss in U because each worker is choosing the amount of work, or the kind of job, which maximizes his private utility rather than the amount of work or kind of job which maximizes social utility. In general, the redistributive gains versus the losses caused by tax/transfer-induced switching is the major tradeoff in the theory of optimal income taxes and welfare payments—both with and without tagging.

III. A Simple Example of Optimal Taxes and Subsidies with Tagging

Section I gave formula (2) which indicated that tagging improved the relation between the marginal tax rate and the minimum subsidy to *tagged* poor people. Loosely, it could be said that tagging will in consequence reduce the cost of income redistribution (since, with lower marginal tax rates, there is a smaller gap between social and private returns from work and therefore less loss of consumer's surplus due to redistribution-caused job switching). As a result, it is only natural that tagging increases the optimal transfers to poor people.

A. *The Rudimentary Mirrlees-Fair Model*

As implied by Mirrlees, there are no interesting easily solved algebraic examples of the optimal income tax with a continuum of abilities. There is no question that tagging, since it adds an additional degree of freedom, makes the problem still harder. Therefore, the example presented here is a much simplified version of the Mirrlees-Fair general case.

The example here is the most rudimentary model in which the optimal tax structure, both with and without tagging, is dictated by the tradeoffs between the deadweight loss due to taxes and subsidies and the gains of redistribution from rich to poor. Instead of a continuum of workers (as in Mirrlees), there are just two types: skilled and unskilled; instead of a continuum of output dependent upon labor input, there are just two types of jobs: difficult jobs (denoted by subscript D) and easy jobs (denoted by subscript E). Instead of a marginal condition describing the optimal tax reflecting continua of both labor input and worker types and the corresponding use of the calculus of variations, the optimum tax is characterized by a binding inequality constraint, which results from the discrete calculus corresponding to the discrete number of job types and worker types.

It is assumed that there are an equal number of skilled and unskilled workers. Skilled workers may work in either difficult or easy jobs, but unskilled workers may work only in easy jobs.[3] The output of a skilled worker in a difficult job is q_D, which is a constant independent of the number of workers in such jobs. Similarly, the output of both skilled and unskilled workers in easy jobs is q_E, which is also a constant independent of the number of workers in such jobs. These data are summarized in Table 2, which gives the technology of the model. Of course, output in difficult jobs exceeds output in easy jobs, so that $q_D > q_E$.

[3]The model works out equivalently if unskilled workers can work in different jobs but have great distaste for the extra effort required.

TABLE 2—OUTPUT OF WORKER BY TYPE OF WORKER BY TYPE OF JOB

Type of Worker (Percent of Workforce)	Type of Job Difficult	Easy
Skilled (50%)	q_D	q_E
Unskilled (50%)	Not applicable	q_E

Note: $q_D > q_E$

The economy is competitive, so that pretax, pretransfer pay in each job is the worker's marginal product in that job. The utility of each worker depends upon after-tax, after-transfer income and upon the nonpecuniary returns of his job. The utility functions can be written as a separable function of the pecuniary and the nonpecuniary returns. Let t_D denote the taxes paid by workers in difficult jobs (with income q_D), and let t_E denote transfers to workers in easy jobs (with income q_E). After-tax income in difficult jobs is $q_D - t_D$; after-transfer income in easy jobs is $q_E + t_E$. The utility of skilled workers in difficult jobs is $u(q_D - t_D) - \delta$, and the utility of both skilled and unskilled workers in easy jobs is $u(q_E + t_E)$. The parameter δ reflects the nonpecuniary distaste of workers for difficult jobs due to the greater effort necessary. Of course, $u' > 0, u'' < 0$. It is further assumed that $u(q_D) - \delta > u(q_E)$; otherwise, easy jobs dominate difficult jobs, so that, at the optimum, all workers (trivially) work in easy jobs without paying taxes or receiving transfers. The preceding data are summarized in Table 3.

In the absence of tagging, the Mirrlees-

TABLE 3—UTILITY OF WORKERS BY TYPE OF WORKER BY TYPE OF JOB, WITH TAXES t_D ON PERSONS WITH PRETAX INCOME q_D, AND TRANSFERS t_E TO PERSONS WITH PRETAX INCOME q_E

Type of Worker (Percent of Workforce)	Type of Job Difficult	Easy
Skilled (50%)	$u(q_D - t_D) - \delta$	$u(q_E + t_E)$
Unskilled (50%)	Not applicable	$u(q_E + t_E)$

Note: $u(q_D) - \delta > u(q_E)$

Fair optimal income tax, as applied to this model, is obtained by choosing a tax on income in difficult jobs t_D and a transfer to income in easy jobs t_E, subject to the constraint that qualified workers will choose skilled or unskilled jobs depending upon which one yields greater utility (after taxes), and also subject to the constraint that taxes equal transfers. In mathematical form this becomes the maximization problem to choose t_D and t_E to maximize U,

$$(6) \quad U = \frac{1}{2} \max\{u(q_D - t_D) - \delta, u(q_E + t_E)\} + \frac{1}{2} u(q_E + t_E)$$

subject to

(7a)

$$t_D = t_E \quad \text{if} \quad u(q_D - t_D) - \delta \geq u(q_E + t_E)$$

(7b)

$$t_E = 0 \quad \text{if} \quad u(q_D - t_D) - \delta < u(q_E + t_E)$$

It is convenient to denote optimal values with an asterisk. Thus the optimal value of U is U^*, of t is t^*, and of t_E is t_E^*.

The maximand (6) consists of the sum of the utilities of skilled and unskilled workers weighted by their respective fractions of the population. The utility of a skilled worker is $\max\{u(q_D - t_D) - \delta, u(q_E + t_E)\}$ since skilled workers are assumed to work in difficult jobs if $u(q_D - t_D) - \delta \geq u(q_E + t_E)$, and in easy jobs otherwise. Equations (7a) and (7b) jointly reflect the balanced budget constraint. If skilled workers work in difficult jobs, the tax collection per skilled worker is t_D. If tax collections equal transfers, $t_D = t_E$ (which is (7a)). However, if skilled workers work in easy jobs, they must receive the same transfer as unskilled workers. As a result, the condition that taxes equal transfers implies that $t_E = 0$, which is (7b).

Tagging does not occur in this maximization, since skilled and unskilled workers alike receive the same transfer t_E if they work in easy jobs.

Two equations, (8) and (9), characterize the optimal tax-*cum*-transfer rates t_D^* and t_E^* which maximize U:

$$(8) \quad t_D^* = t_E^*$$

$$(9) \quad u(q_D - t_D^*) - \delta = u(q_E + t_E^*)$$

Of course, (8) is the tax-equal-transfer balanced budget constraint. Equation (9) expresses the additional condition that, at the optimum, as much is redistributed from skilled to unskilled workers as possible, subject to the constraint that any greater redistribution would cause skilled workers to switch from difficult to easy jobs. (Any increase in t_D above t_D^*, or in t_E above t_E^*, results in a shift of all skilled workers into easy jobs.) As a result of this threatened shift, the deadweight loss due to a marginal increase in taxes or in transfers exceeds the returns from any redistributive gain.[4] Thus, our model, although rudimentary, has an optimal tax-*cum*-transfer schedule which reflects the tradeoffs of Mirrlees-Fair: the optimal tax/transfer policy being determined both by the gains from redistribution and the losses due to labor-supply shifts in response to changes in taxes and transfers.

B. *Tagging Introduced into Rudimentary Mirrlees-Fair Model*

Now consider how tagging will alter the Mirrlees-Fair maximization and its solution. Suppose that a portion β of the unskilled workers can be identified (i.e., tagged) as unskilled and given a tax/transfer schedule different from that of other workers. In the altered model with tagging, let T_D denote the taxes paid by untagged workers in difficult jobs; let T_E denote transfers (perhaps negative) paid to untagged workers in easy jobs; and let τ denote the transfer to tagged workers (all of whom work in easy jobs). Table 4 compares the tax/transfer schedule of the earlier

[4] It also happens in this maximization that any further increase in taxes or in transfers at the margin causes such a large and discontinuous shift in the number of workers earning high incomes in difficult jobs that such an increase also decreases the revenues available for redistribution to unskilled workers.

TABLE 4—TAXES ON DIFFICULT JOBS AND TRANSFERS TO EASY JOBS IN MODELS WITH AND WITHOUT TAGGING

	Model without Tagging	Model with Tagging
Tax on Difficult Job	t_D	T_D
Transfer to Easy Job (workers untagged)	t_E	T_E
Transfer to Easy Job (workers tagged)	Not Applicable	τ

model without tagging and the tax schedule of the current model with tagging.

Using Table 4, it is easy to construct Table 5, which gives the utility of workers by type of job after taxes and after transfers. Table 5 differs from Table 3 by addition of the bottom row, which represents the utility of tagged workers in easy jobs who receive the transfer τ.

Using the data in Table 5, it is easy to see that, with tagging, the optimum tax-*cum*-transfer policy is to choose the values (T_D, T_E, τ) that maximize U^{Tag}, where:

$$(10)\quad U^{Tag} = \frac{1}{2}\max\{u(q_D - T_D) - \delta, u(q_E + T_E)\} + \frac{1}{2}(1-\beta)u(q_E + T_E) + \frac{1}{2}\beta u(q_E + \tau)$$

subject to the balanced budget constraints (11a) and (11b):

$$(11a)\quad T_D = (1-\beta)T_E + \beta\tau \quad \text{if } u(q_D - T_D) - \delta \geq u(q_E + T_E)$$

$$(11b)\quad (2-\beta)T_E + \beta\tau = 0 \quad \text{if } u(q_D - T_D) - \delta < u(q_E + T_E)$$

Again, denote the optimum values with an asterisk: T_D^*, T_E^*, τ^*, and U^{Tag*}.

The maximand U^{Tag} is the sum of the utility of all three types of workers—skilled, untagged unskilled, and tagged unskilled weighted by their respective fractions of the population. The utility of skilled workers is $u(q_D - T_D) - \delta$ or $u(q_E + T_E)$, dependent upon whether they choose difficult or easy jobs. Equations (11a) and (11b) are the tax-equal-transfer, balanced-budget constraints.

TABLE 5—UTILITY OF WORKER BY TYPE OF WORKER BY TYPE OF JOB WITH TAGGING: UNTAGGED WORKERS PAY TAXES T_D IN DIFFICULT JOBS AND RECEIVE TRANSFERS T_E IN UNSKILLED JOBS; TAGGED WORKERS RECEIVE A TRANSFER τ IN UNSKILLED JOBS

Type of Worker (Fraction of Workforce)	Type of Job: Difficult	Easy
Skilled (Untagged) (1/2)	$u(q_D - T_D) - \delta$	$u(q_E + T_E)$
Unskilled (Untagged) $((1-\beta)/2)$	Not Applicable	$u(q_E + T_E)$
Unskilled (Tagged) $(\beta/2)$	Not Applicable	$u(q_E + \tau)$

The respective equation applies accordingly as skilled workers are in difficult or in easy jobs.

In the Appendix, it is shown that with $u(q_D) - \delta > u(q_E)$, for $0 < \beta \leq 1$, the optimal transfer to tagged workers τ^* exceeds the optimal transfer to untagged unskilled workers t_E^* in the model without tagging. With $\beta = 1$, complete equality of income is attained at the optimum. In this precise sense, tagging increases the optimum transfers to those who are identified as poor and given special tax treatment.

The difference between the tagging and the nontagging optimization is clear: with tagging, for a given increased subsidy to tagged people, there is a smaller decline in the income differential between difficult and easy work, since T_E need not shift, and there is therefore a smaller tendency for workers to shift from difficult to easy jobs with a given redistribution of income. As a result, optimal transfers to tagged workers are greater with tagging than in its absence.

An outline of the proof, which is given in the Appendix, illustrates the application of this logic more particularly. The proof shows that, at the optimum, the rate of taxation of workers in difficult jobs and the rate of transfer to untagged workers in easy jobs is taken up to the point that any further increase in either of those two rates will

induce skilled workers to shift into easy jobs. This is reflected by the optimization condition (12), which is exactly analogous to the similar optimization condition (9) in the untagged case:

$$u(q_D - T_D^*) - \delta = u(q_E + T_E^*) \tag{12}$$

It is then shown by contradiction that τ^* (the optimal transfer to tagged workers) exceeds T_E^* (the optimal transfer to unskilled untagged workers). Suppose the contrary (i.e., $\tau^* \leq T_E^*$). In that case, a marginal decrease in T_E and a marginal increase in equal dollar amount in τ can cause no decrease in utility, while it allows some additional redistribution to be made from skilled workers in difficult jobs to other workers without inducing any skilled workers to switch from difficult into easy jobs. Since total utility U^{tag} is sure to be increased by at least one of these two changes and not decreased by the other, the optimality of τ^* and T_E^* is contradicted. At the optimum, therefore, τ^* must be greater than T_E^*.

Knowing that $\tau^* > T_E^*$, as has been shown, knowing that T_D^* and T_E^* satisfy (12), and knowing that t_D^* and t_E^* satisfy the similar condition (9), $u(q_D - t_D^*) - \delta = u(q_E + t_E^*)$, the budget constraints can be used to show that $\tau^* > t_E^*$.

IV. Generalized Problem

In the example in the last section, there was no opportunity for people to change the characteristics by which they were tagged. Age, race, and sex are real life examples of such characteristics. However, there are also redistribution programs in which people, by some effort or with some loss of utility, may alter their characteristics, thereby becoming members of a tagged group. The most commonly cited example of this concerns families who allegedly have separated in order to obtain payments under the Aid to Dependent Children program (see Daniel Moynihan).

To consider the case more generally, in which group membership is endogenous, this section presents a general model. It then becomes an empirical (rather than a theoretical) question to determine what amount of tagging (and quite possibly the answer is none) will maximize aggregate utility U. There is no major theorem in general, unless it is the falsity of the proposition to which the previous section gave a counterexample, that a uniform negative income tax is always superior to a welfare system that gives special aid to people with special problems or characteristics.

In general, we may assume the goal is to choose functions $t_\gamma(y_\gamma)$ to maximize

$$U = \int u_x f(x)dx \tag{13}$$

where $f(x)$ denotes the distribution of people of type x, and where the utility of such a person depends on his after-tax income, his characteristics, and the group to which he belongs γ, or

$$u_x = u(y - t, x, \gamma) \tag{14}$$

In the real world, of course, tagging is not costless, one of the major complaints against the current welfare system being its cost of administration. Let Γ be the grouping of people into various subgroups of the population, and let $c(\Gamma)$ be the administrative cost of such tagging.

U is maximized subject to two constraints, the first being that taxes equal transfers plus administrative costs, or

$$\int_x t_\gamma(y(x), \gamma(x)) f(x)dx + c(\Gamma) = 0 \tag{15}$$

where $\gamma(x)$ is the group to which an individual of type x belongs, and the second being that an individual of type x chooses his labor input and the group to which he belongs to maximize

$$u[w(x,\gamma)L(x,\gamma) - t_\gamma(w(x,\gamma)L(x,\gamma)), x, \gamma] \tag{16}$$

where $w(x,\gamma)$ is the wage of a person of characteristic x belonging to group γ, and $L(x,\gamma)$ is the labor input.

In sum, this is the generalization of Mirrlees' (and Fair's) problem to taxation with tagging. I have taken the trouble to

specify this general problem since it is important to note the potential endogeneity of the tagged characteristics and of administrative costs.

V. Cost-Benefit Evaluation of Manpower Programs and Tagging

Another type of program in which tagging is important is manpower training programs. Typically, such programs in the United States have aimed at improving the skills of the disadvantaged and the temporarily unemployed. Because of formal eligibility requirements, and also because of the self-selectivity of the trainees, people in special need are identified (or tagged) by such programs.

There has been an intensive effort in the United States to evaluate the benefits and costs of such programs, so much so that there have been extensive "reviews of the reviews" (see David O'Neill). The studies have typically (but with some exceptions) found that the benefits of manpower training programs, as conventionally accounted, have been less than the costs. But because of the value of tagging done by such programs, a benefit-cost ratio of less than unity is not sufficient reason for their curtailment.

This last point can be made formally in terms of the tagging models in Sections III and IV. A manpower program could be introduced into the model in Section III by assuming that, at a given cost per worker, an unskilled worker who is previously untagged can be made into a skilled worker. The costs of such a program, as usually accounted, are its costs of operation plus the wages foregone by workers while engaged in training. The cost of operation becomes an additional term in the balanced budget constraint (analogous to the term $c(\Gamma)$ in (15)). The benefits from the program are the increase in the pretax, pretransfer wages of the worker subsequent to training. It is easy to construct an example in which the benefits (thus accounted) are less than the costs (thus accounted), yet U^{Tag} is greater with the program than in its absence, because the program tags unskilled workers and makes income redistribution possible with relatively little distortion to the incentive structure.

An unrigorous calculation using consumer's surplus logic shows that the tagging benefits of manpower programs may be substantial. Consider two subgroups of the population, both of which are young and both of which have low current incomes. One group is skilled but has low current income because it is building up human capital; the other group is unskilled and has low current income for that reason; it also has low permanent income.

Let there be a manpower training program. At a cost of c dollars, the permanent income of a young unskilled worker can be raised by \$1. The costs of this program (as usually accounted) are c dollars, and its benefits are \$1. Considering consumer's surplus and assuming that there is a deadweight loss of λ per dollar due to taxes to pay for the program, the cost of the program, inclusive of deadweight loss is $c(1 + \lambda)$.

Now compare the advantages of this training program to a negative income tax that gives lump sum transfers to all young workers, whether skilled or unskilled. Let unskilled workers be a fraction θ of the total population. To redistribute \$1 to an unskilled young worker, a total of $1/\theta$ dollars must be redistributed to all young people.

Which scheme—the manpower training program or the negative income tax—is the cheaper way of redistributing \$1 to unskilled workers? The cost, inclusive of deadweight loss of the manpower program, is $c(1 + \lambda)$. The cost, inclusive of deadweight loss of the negative income tax, is the deadweight loss on $1/\theta$ dollars, plus the \$1 redistributed, or $\lambda/\theta + 1$. Which scheme is cheaper depends upon whether $c(1 + \lambda)$ is greater or less than $(\lambda/\theta + 1)$.

Let λ be .05 and let θ be .1, numbers which are not unrepresentative of reasonable parameters for deadweight loss due to income taxation and the fraction of the population eligible for a typical manpower training program such as the Job Corps. If the benefit-cost ratio of the manpower program $(1/c)$ is less than .7, the negative income tax is the cheaper method of redis-

tribution; if the benefit-cost ratio is greater than .7, the manpower program is preferable.

VI. Summary and Conclusions

This paper has identified the important tradeoffs in the design of institutions to redistribute income. Some types of programs, either by their eligibility requirements or by the self-selection of the beneficiaries, identify (tag) people who are in special need. With tagging, taxpayers (as opposed to beneficiaries) are denied the benefit of the transfer, so that in effect a *lump sum* transfer is made to tagged people.

In contrast, with a negative income tax, a grant is made to all taxpayers and this grant must be recovered to achieve the same net revenue. This recovery results in high marginal tax rates, whose disincentive effects are the major disadvantage of a negative income tax. This disadvantage, however, must be weighed against the disadvantages of tagging, which are the perverse incentives to people to be identified as needy (to be tagged), the inequity of such a system, and its cost of administration.

The problem of the optimal redistributional system, both with and without tagging, has been set up in the framework of the Mirrlees-Fair optimal income tax. It was shown in a special example that if a portion of the poor population could be identified (costlessly, in this example), total welfare U could be raised by giving increased subsidies to the tagged poor.

Finally, the consequences of tagging for manpower programs were discussed. Since tagging is a benefit of most manpower programs, benefit-cost ratios need not exceed unity to justify their existence. In fact, an example showed that benefit/cost ratios could be significantly less than one (.7 in the example), and a manpower program might still be preferable to a negative income tax as a method of income redistribution.

APPENDIX

THEOREM 1: *Using the definitions of* τ^* *and* t_E^* *in Section III, and also the models in that section, if* $u(q_D) - \delta > u(q_E)$ *and* $0 < \beta \leq 1, \tau^* > t_E^*$.

PROOF:

The proof proceeds by five propositions. Propositions 1 and 2 make variational arguments which show that at the maximum as much must be redistributed from skilled workers as possible without inducing them to switch into easy jobs. This yields the condition:

$$u(q_D - T_D^*) - \delta = u(q_E + T_E^*) \tag{A1}$$

It is similarly true without tagging that

$$u(q_D - t_D^*) - \delta = u(q_E + t_E^*) \tag{A2}$$

From (A1) and (A2) it can be easily shown (Proposition 3) that if $T_D^* > t_D^*$, $T_E^* < t_E^*$ (and vice versa).

Proposition 4 then shows that $\tau^* \geq t_E^*$. There are two cases. In one case, $T_D^* < t_D^*$. If $T_D^* < t_D^*$, by Proposition 3, $T_E^* > t_E^*$. Suppose $t_E^* \geq \tau^*$. A variational argument shows that this cannot be a maximum, for a decrease in T_E^* and an increase in τ^* can increase U^{Tag}. In the other case, $T_D^* \geq t_D^*$. But if $T_D^* \geq t_D^*$, by Proposition 3, $T_E^* \leq t_E^*$. It follows from the balanced budget constraints that if T_E^* is smaller than t_E^*, but also, T_D^* is larger than t_D^*, that τ^* must be larger than t_E^*. As a result, in both Case I and Case II, $\tau^* \geq t_E^*$. Proposition 5 shows that the inequality is strict.

PROPOSITION 1: $u(q_D - T_D^*) - \delta \geq u(q_E + T_E^*)$

PROOF:

Suppose otherwise. Then,

$$U^{Tag^*} = \frac{1}{2}\{(2-\beta)u(q_E + T_E^*) + \beta u(q_E + \tau^*)\} \leq u(q_E) \tag{A3}$$

by the concavity of u and the constraint (11b) that $(2-\beta)T_E^* = -\beta\tau^*$. Since $u(q_D) - \delta > u(q_E)$ by assumption,

$$u(q_E) < \frac{1}{2}\{u(q_D) - \delta + u(q_E)\} \tag{A4}$$

Since $T_D = T_E = \tau = 0$ is a feasible tax/transfer vector (satisfying budget constraint

(11)), and with

(A5) $U^{Tag} = \frac{1}{2}\{u(q_D) - \delta + u(q_E)\}$

the optimality of U^{Tag*} is contradicted by (A3), (A4), and (A5). By this contradiction,

(A6) $u(q_D - T_D^*) - \delta \geq u(q_E + T_E^*)$

PROPOSITION 2:

(A7) $u(q_D - T_D^*) - \delta = u(q_E + T_E^*)$

PROOF:

Suppose that $u(q_D - T_D^*) - \delta > u(q_E + T_E^*)$. A variational argument shows that (T_D^*, T_E^*, τ^*) is not optimal.

Let $T'_D = T_D^* + \epsilon$

$T'_E = T_E^* + \epsilon/(1-\beta)$

(A8) $U^{Tag}(T'_D, T'_E, \tau^*) = U^{Tag}(T_D^*, T_E^*, \tau^*) + \epsilon/2[-u'(q_D - T_D^*) + u'(q_E + T_E^*)] + \sigma^2(\epsilon)$

where $o^2(\epsilon)$ is an expression with $\lim_{\epsilon \to 0} o^2(\epsilon)/\epsilon = 0$. But since $u(q_D - T_D^*) - \delta > u(q_E + T_E^*)$ by assumption,

(A9) $u'(q_D - T_D^*) < u'(q_E + T_E^*)$

by the concavity of u.

Therefore, by (A8), $U^{Tag}(T'_D, T'_E, \tau^*) > U^{Tag}(T_D^*, T_E^*, \tau^*)$ for ϵ sufficiently small, which contradicts the optimality of (T_D^*, T_E^*, τ^*). Therefore, $u(q_D - T_D^*) - \delta \leq u(q_E + T_E^*)$.

By Proposition 1, $u(q_D - T_D^*) - \delta \geq u(q_E + T_E^*)$. Therefore,

(A10) $u(q_D - T_D^*) - \delta = u(q_E + T_E^*)$

PROPOSITION 3: $T_D^* > t_D^*$ *if and only if* $T_E^* < t_E^*$

PROOF:

Suppose $T_D^* > t_D^*$. By Proposition 2

(A11) $u(q_D - T_D^*) - \delta = u(q_E + T_E^*)$

By similar logic,

(A12) $u(q_D - t_D^*) - \delta = u(q_E + t_E^*)$

If $T_D^* > t_D^*$, then

(A13) $u(q_D - T_D^*) < u(q_D - t_D^*)$

whence

(A14) $u(q_E + T_E^*) = u(q_D - T_D^*) - \delta < u(q_D - t_D^*) - \delta = u(q_E + t_E^*)$

(A15) $T_E^* < t_E^*$

Similarly, if $T_D^* < t_D^*$, $T_E^* > t_E^*$

PROPOSITION 4: $\tau^* \geq t_E^*$

PROOF:

Suppose

(A16) $\tau^* < t_E^*$

It will be shown that the optimality of τ^* or of t_E^* is contradicted. Two cases will be analyzed:

Case I: $T_D^* < t_D^*$
Case II: $T_D^* \geq t_D^*$

Case I: By Proposition 3, if $T_D^* < t_D^*$,

(A17) $T_E^* > t_E^*$

But then

(A18) $U^{Tag}(T_D^*, T_E^* - \epsilon, \tau^* + (1-\beta)/\beta\epsilon) = U^{Tag}(T_D^*, T_E^*, \tau^*) - (1-\beta)\epsilon/2u'(q_E + T_E^*) + \beta\frac{1-\beta}{\beta}\epsilon/2u'(q_E + \tau^*) + o^2(\epsilon)$

which last equation (A18) for sufficiently small ϵ

(A19) $> U^{Tag}(T_D^*, T_E^*, \tau^*)$

since $u'(q_E + T_E^*) < u'(q_E + t_E^*) < u'(q_E + \tau^*)$ by the concavity of u and by both the inequality (A17), $(T_E^* > t_E^*)$, and the supposition (A16), $(t_E^* > \tau^*)$. The inequality (A19) contradicts the optimality of (T_D^*, T_E^*, τ^*). Therefore, if $T_D^* < t_D^*$, $\tau^* \geq t_E^*$.

Case II: $T_D^* \geq t_D^*$.

Suppose again

(A20) $\tau^* < t_E^*$

We will show a contradiction. By Proposition 3, if $T_D^* \geq t_D^*$,

(A21) $T_E^* \leq t_E^*$

By inequality (A21), ($T_E^* \leq t_E^*$), the budget constraint (7a), ($t_D^* = t_E^*$), and inequality (A20), ($\tau^* < t_E^*$),

$$(A22) \quad T_D^* \geq t_D^* = t_E^* > (1 - \beta)T_E^* + \beta\tau^*$$

which contradicts the budget constraint (11a), which states:

$$(A23) \quad T_D^* = (1 - \beta)T_E^* + \beta\tau^*$$

Hence, if $T_D^* \geq t_D^*$, $\tau^* \geq t_E^*$.

Combining Cases I and II, it has been shown that $\tau^* \geq t_E^*$.

PROPOSITION 5: $\tau^* > t_E^*$

PROOF:

It remains to show that $\tau^* \neq t_E^*$. Suppose the contrary, that $\tau^* = t_E^*$. A contradiction will be demonstrated. By Proposition 3 at the optimum

$$(A24) \quad u(q_D - T_D^*) - \delta = u(q_E + T_E^*)$$

and similarly,

$$(A25) \quad u(q_D - t_D^*) - \delta = u(q_E + t_E^*)$$

The optimum (T_D^*, T_E^*, τ^*) and (t_D^*, t_E^*) must also satisfy the budget constraints (7a) and (11a):

$$(A26) \quad T_D^* = (1 - \beta)T_E^* + \beta\tau^*$$

$$(A27) \quad t_D^* = t_E^*$$

Add to the system (A24) to (A27) the assumption (A28):

$$(A28) \quad \tau^* = t_E^*$$

An optimum with $\tau^* = t_E^*$ must satisfy the five relations (A24) to (A28). These five equations constitute a system of five equations in the five variables ($T_D^*, T_E^*, \tau^*, t_D^*, t_E^*$), with unique solution with the property

$$T_D^* = T_E^* = \tau^* = t_D^* = t_E^*$$

Let

$$(A29) \quad T'_D = T_D^* + 2\epsilon_1$$

$$(A30) \quad T'_E = T_E^* - 2\epsilon_2$$

$$(A31) \quad \tau' = \tau^* + \frac{1 - \beta}{\beta} 2\epsilon_2 + \frac{1}{\beta} 2\epsilon_1$$

with

$$(A32) \quad \epsilon_1 < \frac{u'(q_D - T_D^*)}{u'(q_E + T_E^*)} \epsilon_2$$

Then,

$$(A33) \quad U^{Tag}(T'_D, T'_E, \tau') = U^{Tag}(T_D^*, T_E^*, \tau^*) - \epsilon_1 u'(q_D - T_D^*) - (1 - \beta)\epsilon_2 u'(q_E + T_E^*) + \beta \frac{\epsilon_1}{\beta} u'(q_E + \tau^*) + \beta \frac{1 - \beta}{\beta} \epsilon_2 u'(q_E + \tau^*) + o^2(\epsilon_1) + o^2(\epsilon_2)$$

Since $\tau^* = T_E^*$, for (ϵ_1, ϵ_2) sufficiently small $U^{Tag}(T'_D, T'_E, \tau') > U^{Tag}(T_D^*, T_E^*, \tau^*)$, which contradicts the optimality of (T_D^*, T_E^*, τ^*). Hence, $\tau^* \neq t_E^*$. And, using Proposition 4, $\tau^* > t_E^*$.

REFERENCES

R. C. Fair, "The Optimal Distribution of Income," *Quart. J. Econ.*, Nov. 1971, *85*, 557–79.

J. A. Mirrlees, "An Exploration in the Optimal Theory of Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 175–208.

D. P. Moynihan, "The Negro Family: The Case for National Action," in L. Rainwater and W. L. Yancey, eds., *The Moynihan Report and the Politics of Controversy*, Cambridge, Mass. 1967.

D. M. O'Neill, "The Federal Government and Manpower: A Critical Look at the MDTA-Institutional and Job Corps Programs," American Enterprise Institute for Policy Research, 1973.

New York Times, Aug. 9, 1969; Apr. 21, 1970; July 23, 1971.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1974.

Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance, and Law Enforcement

By MILTON HARRIS AND ARTUR RAVIV*

When decision-making authority is delegated from one agent to another, contractual arrangements are often used to allocate resources and outputs. Such situations may be analyzed using the theory of principal-agent relationships. This theory seeks to characterize optimal contracts and explain observed arrangements. Examples which fit the "agency paradigm" include employer-employee, insurer-insured, and owner-manager relations. In this paper, we report some results which significantly extend the theory of agency to situations characterized by a divergence of incentives between the two parties and asymmetric information with opportunities for acquiring information. In addition, we discuss several applications of this theory.

The theory of optimal contracts under conditions of uncertainty has received considerable attention. Kenneth Arrow and Robert Wilson (1968) were concerned with the optimal sharing of purely exogenous risk. Wilson (1969) and Stephen Ross considered situations in which the risk could be affected by the actions of the agents. They analyze contracts which induce similar attitudes toward risk on the part of the agents, thus allowing the possibility that decentralized decision making will be optimal. Conditions under which such arrangements are indeed Pareto optimal are also investigated. In their models, incentive problems arise purely as a consequence of diverse attitudes toward risk among the agents. A. Michael Spence and Richard Zeckhauser, in the context of insurance contracts, introduced the problem of a divergence in incentives due to the action of an agent together with differential information among agents. More recently, Joseph Stiglitz (1975a) analyzed incentive contracts between employers and employees. With regard to differential information, both Spence-Zeckhauser and Stiglitz assume one of two extreme cases: either the agent's action is known by everyone with certainty (in which case there is no differential information), or no information about the agent's action is available to anyone except the agent himself. A somewhat intermediate case was analyzed by Robert Townsend in which the exact information possessed by one agent can be made available to the other at some cost.

Two important aspects of agency relationships are not fully explored in the literature on the theory of contracts. First, most agency relationships must deal with incentive problems which arise because the agent would prefer to work less, other things equal, while the principal is indifferent to the level of the agent's effort, other things (i.e., his share of the payoff) equal. This type of incentive problem is somewhat different from the one considered by Wilson (1969) and Ross in which a divergence in incentives results only from different attitudes toward risk. Second, in most instances an agent may acquire information about other agents' actions. This possibility was discussed in interesting papers by Armen Alchian and Harold Demsetz and by C. Michael Jensen and William Meckling. The quality of the information obtained through monitoring (or supervising) depends on the resources committed to this activity as well as on the available monitoring technology. Furthermore, as Stiglitz

*Carnegie-Mellon University. We would like to acknowledge helpful discussions with Ed Prescott and Rob Townsend, as well as suggestions of Stephen Ross, Martin Hellwig, an anonymous referee, and the managing editor of this *Review*.

points out, "... the amount (or quality) of supervision will affect both the optimal incentive scheme which will be used and the level of expected utility which the individual will attain" (1975a, p. 572). Consequently, the optimal incentive contract will depend on the available monitoring technology. In this paper, we explore these aspects of the agency problem.

Our analysis is based on a model in which there are two individuals: one, denoted the agent, takes an action which together with the realization of an exogenous random variable results in the payoff to be divided between the agent and the other individual, denoted the principal. Incentive problems arise because the agent has a disutility for the action while the principal does not. We distinguish two versions of this model. In the first, the agent is assumed to take his action without any information regarding the realization of the exogenous random state. In the second version, the agent is assumed to know the value of the random state before taking his action. In both versions, the object of the analysis is to discover the form of the Pareto optimal contract, that is, how the optimal sharing arrangement for the payoff depends on the observed variables. In particular, our analysis deals with the following issues: When would we expect to observe performance-contingent contracts, and what would be the form of such contracts; when performance is not observable, under what conditions would we expect contracts to depend on imperfect estimates of performance, what types of estimators would be used, and how would they be incorporated into the contract.

Our results are discussed in terms of the employer-employee relationship. They are, however, more general, and we discuss their implications for three other agency relationships.

The first application is to the analysis of employment contracts based on training or education and ability. Here we show under what conditions and in what form it is optimal to make employment contracts contingent on training or ability. In particular we show when the type of contract used in the "signaling literature" (see Spence, 1973, 1974, 1976; John Riley; Stiglitz, 1975b) is Pareto optimal.

The second application is to health insurance contracts. We show when indemnity insurance, that is, contracts in which the insurance payment depends only on the degree of illness, is optimal. We also show when the optimal policy provides payment contingent on the level of medical care chosen. The case in which the level of medical care chosen is not directly observable is also considered.

The third application of the analysis is to the problem of procuring the optimal amount of law enforcement, an issue which was addressed by Gary Becker and George Stigler. We exhibit conditions on the information structure under which the standard type of compensation arrangement for the enforcer (i.e., a salary) would lead to an inefficiently large degree of malfeasance or nonfeasance (shirking). We also show under what conditions these inefficiencies could be resolved by the use of contracts which we exhibit. In particular, we provide conditions under which some suggestions of Becker and Stigler would be optimal.

I. The Agency Model

In this section we describe the model with which we address the issues mentioned above (a more detailed and formal treatment may be found in the authors' working paper). In this model there are two individuals, "the principal" and "the agent." For concreteness, in describing the model and results, we refer to these individuals as the employer and worker, respectively.

The worker chooses a level of effort (or action) a, which together with the realization of some exogenous random variable θ determines the value of the worker's product x (the payoff). The random variable θ may be interpreted as the result of any exogenous uncertain event which affects the worker's productivity, for example, the weather, equipment failure, the price of the product, etc. We represent the relationship

among the value of the worker's output, the worker's effort, and the realization of the random variable by a production function X, that is,

$$x = X(a, \theta) \tag{1}$$

We assume that greater effort by the worker results in greater output for any value of θ (i.e., $X_1 > 0$ for all a, θ).

Two important cases may be distinguished regarding the information available to the worker when he decides upon his level of effort. First, in Model 1, we assume that he does *not* know the value of θ when he chooses his effort. For example, a sharecropper may not know what the weather will be when he plants his crop; a lawyer, when he prepares his case, may not know which judge will preside. Second, in Model 2, we assume that the worker *does* observe the value of θ and chooses his effort contingent on this observation. For example, a salesman may observe the state of demand before deciding how many calls to make.

The division of the product will be determined by a function which may depend on any variable which is observable by both parties, that is, whose value is known by both parties when the product is divided. This assumption rules out contracts which are not incentive compatible. The function and its list of arguments denoted $(S; z)$ will be called a *contract*. The value of $S(z)$ is interpreted as the worker's share of the product while the employer's share is $x - S(z)$. The product x is assumed always to be observable by both parties. In addition to x, the list z might include the effort a, provided, of course, that it is observable. If a is not observable, z might include an estimator, denoted α of a. For example, an estimator of the sharecropper's effort might be the amount of time he spends in the field. An estimator of the salesman's effort might be the number of miles he drives. When available, the estimator will be called a *monitor*, that is, a monitor is a random variable whose distribution is conditional on a. The class of available monitors is referred to as the *monitoring technology*.

Associated with the worker is a utility function U, whose first argument is the worker's share of the product and whose second argument is his effort. The worker is assumed to prefer less effort to more effort, other things equal. We therefore assume that

$$U_1 > 0; \qquad U_2 < 0 \tag{2}$$

We will often assume that the worker is risk averse, that is, that U is strictly concave in the first argument.

Given a contract $(S; z)$, in Model 1 the worker determines his effort by solving the following maximization problem:

$$\max_a E_\theta U[S(z), a] \tag{3}$$

In Model 2, the worker chooses a as a function of θ to solve

$$\max_a U[S(z), a] \tag{3'}$$

In both models, the effort chosen will depend on the functional form of S; in Model 2, it will also depend on the realization of the exogenous random variable.

Associated with the employer is a utility function V, which is a function *only* of his share of the product. We assume V is monotone increasing, concave (i.e., the employer is either risk neutral or risk averse). Given a contract $(S; z)$, and a level of effort a chosen by the worker according to either (3) or (3'), the employer's utility is

$$E_\theta V[X(a, \theta) - S(z)] \tag{4}$$

This concludes our presentation of the model. We turn in the next section to a description of our results.

II. Results

Our results characterize the Pareto optimal contracts under various assumptions concerning the availability of information. In particular, optimal contracts are investigated under alternative assumptions on the observability of the worker's effort and the random variable and on the existing tech-

nology for monitoring the worker's effort. We are interested in Pareto optimal contracts on the supposition that observed contracts will have the property that, given the availability of information, neither agent's expected utility can be increased without decreasing the expected utility of the other agent. Thus we seek to characterize the contracts which we expect to arise under various information structures. Our results are simply discussed here without formal derivations; these may be found in our earlier paper.

To begin, suppose that the realization of the random variable (also referred to as "the state of nature" or simply "the state") is freely observable by both employer and worker when the product is distributed. In this case, our first result implies that making the contract contingent on the worker's effort provides no gains over contracts which depend only on output and the observed value of θ. Therefore, if the state is freely observable, we would *not* expect to observe contracts contingent on worker effort. Thus *ex post* uncertainty as to the relationship between the effort and the product is essential if contracts based on worker performance are to be observed. For future reference we state this result as

PROPOSITION 1: *The expected utilities achieved under any contract which depends on the product, the effort, and the state can also be achieved under a contract which depends on the product and the state, but not on the effort. This result holds for both Models* 1 *and* 2.

The above result does not yield information as to the form of the optimal contract when the state is freely observable. This information is provided by

PROPOSITION 2: *The Pareto optimal contract which depends on the product and the state specifies a "standard" output contingent on the state. The worker receives an amount which depends on the standard output and perhaps on the state, plus the difference between the actual output and the standard. The employer receives an amount which depends only on the state and is therefore unaffected by the worker's level of effort. This result holds for both Models* 1 *and* 2.

To illustrate this result, consider the sharecropper example mentioned above. In this case, Propositions 1 and 2 imply that if the weather is the only exogenous random factor affecting the crop, we would expect to observe an arrangement in which the tenant "pays" the landlord an amount of output contingent only on the weather and keeps the remainder of the crop.

Note that Propositions 1 and 2 hold regardless of attitudes toward risk of the employer (landlord) and worker (tenant). Thus the determination of the variables on which the contract will depend (if all are observable) is independent of attitudes toward risk. For example, if output, effort, and the weather are all observable, the contract will depend only on output and weather. This result does depend, however, on the ability of both parties to agree *ex ante* on which types of weather are possible, what the probability attached to each type is, and how weather affects output. It also depends on the ability of both parties to observe *ex post* which type of weather occurred. Attitudes toward risk do play a role in determining the sharing function. For example, if the landlord is risk neutral, the contract will be such that if the tenant puts in sufficient effort to produce the (weather-contingent) standard output, then the tenant's share will be independent of the weather. If the tenant does not render the effort implicit in the standard output, however, his share may be weather-dependent. The landlord's share (i.e., rent) *will* depend on the weather, in general, if the tenant is risk averse.

Attitudes toward risk also play an important role in determining the variables on which the optimal contract will depend when the state is not observable, *ex post*. The next result states that risk aversion on the part of the worker is a necessary and

sufficient condition for contracts which depend explicitly on effort to be superior to contracts which depend only on the output. Intuitively, if the worker is risk neutral, it will be optimal for him to bear all the risk associated with uncertain productivity. In this case, all effects of the worker's performance are internalized, and thus incentive problems are resolved without the use of performance-contingent contracts. When the worker is risk averse, optimality requires some sharing of risk, and therefore the employer bears some of the consequences of the worker's choice of effort. In this case, performance-contingent contracts are superior to contracts based only on output.

PROPOSITION 3: (i) *If the worker is risk neutral, any contract which depends only on output and effort can be dominated (in the Pareto sense) by a contract which depends only on output.*

(ii) *If the worker is risk averse, any contract which depends only on output can be strictly dominated by a contract which depends on both output and effort.*

(iii) *If the worker is risk averse, any contract which depends only on output can be strictly dominated by a contract which depends only on output and the state and is of the form given in Proposition 2.*

These results hold for Models 1 and 2.[1] Next we characterize the form of Pareto optimal contracts which depend on output and effort.

PROPOSITION 4: *Any Pareto optimal contract in the class of contracts depending only on output and effort has the property that the worker's share depends only on the output, provided his effort meets a prespecified criterion. If not, he receives nothing. This result holds for Model 1.*

Thus when the worker's effort is freely observable, and the worker is risk averse, we would expect to observe contracts which stipulate a particular choice of effort by the worker. Since the worker will always choose to meet the requirement, we refer to this contract as a forcing contract.[2]

With respect to the sharecropping paradigm, these results imply that if the tenant is risk averse, and output and effort, *but not the weather,* are observable *ex post,*[3] we would expect the sharing arrangement to include a stipulation of how much effort the tenant is expected to expend. If the tenant puts in the required effort (and he will if his effort is perfectly observable), he will receive a share of the product. This share will depend on the product and the attitudes toward risk of the landlord and tenant. For example if the landlord is risk neutral (and not the tenant), the tenant will get a fixed wage independent of output (or the weather). If, on the other hand, the tenant (and not the landlord) is risk neutral, contracts stipulating the tenant's effort are unnecessary; we expect to observe the landlord receiving a fixed rent with the tenant keeping the residual.

We now turn to some results regarding the use of imperfect estimators of effort as a contingency affecting the distribution of the product. Clearly we would not expect to observe the use of imperfect monitors of effort when there are no gains to using actual effort. We have exhibited above two situations in which there are no gains to acquiring information about the worker's effort (even if such information is available costlessly). Moreover we have shown that in all other cases there are gains to acquiring information, that is, if information of "sufficient quality" can be obtained at a "sufficiently low price," then both individuals can be made better off. The two conditions

[1]For Model 2 the utility function of the worker is assumed to be separable and the production function is one-to-one in the state.

[2]This terminology was suggested to us by Ross.

[3]With respect to the sharecropping example, one might question the assumption that the tenant's effort is observable while the weather is not. We agree that these assumptions are not particularly appropriate to sharecropping. Our model is, however, applicable to a large class of situations, and for many of these, the assumptions are appropriate (see Section III). We use the sharecropping example here only for the purpose of illustrating all our results with a consistent example.

under which there are no gains to monitoring the worker's effort are (a) when the realization of θ, the exogenous random variable affecting output, is observable, and (b) when the worker is risk neutral. From a methodological point of view, these results imply that it is not possible to simplify the analysis of monitoring by assuming away either exogenous uncertainty or risk aversion on the part of the worker.

In general, even when there are gains to perfect information on effort, these gains may be impossible to realize through imperfect monitoring. The introduction of imperfect information on the worker's effort into a contract produces two opposing effects on the welfare of the parties to the contract. First, since the information is imperfect, additional uncertainty is introduced. Since both the employer and worker are risk averse, this additional uncertainty tends to reduce welfare. Second, inclusion of monitoring can motivate the worker to choose a level of effort which, neglecting the first effect, would make both parties better off.

Because the minimum (necessary) conditions for monitoring to be valuable appear to be very difficult to formulate, we explore several sets of sufficient conditions. These results are derived using forcing type contracts. This type of contract specifies that the worker is paid only if the monitoring reveals his effort to be "acceptable," and the size of the payment does not depend on the results of monitoring in any other way. The forcing contract thus requires that a decision be made based on the realization of the monitor. This statistical decision problem is the test of the following hypothesis: the worker's level of effort is one of a set of acceptable levels. The precision of the monitor, in this case, is summarized by the probabilities of type I and type II errors. The conditions on the monitoring technology referred to below are assumptions as to the availability of monitors with various probabilities of type I and type II errors.

Our results regarding the use of contracts based on imperfect monitoring of worker effort can be summarized as follows:

i) First, we establish general sufficient conditions under which the potential gains to monitoring may be realized.

ii) Under additional assumptions on the monitoring technology, we show that any Pareto optimal contract can be approximated to any degree of precision by a forcing type of contract.

iii) Under another set of assumptions on the monitoring technology and the worker's utility function, we show that all the results achievable under perfect information can be obtained even when information is imperfect, using a forcing contract.

The above results are obtained only for Model I.

III. Applications

In this section, three applications of the results of the previous section are analyzed. These results are used to describe the Pareto optimal contractual arrangements, and therefore the contracts we would expect to observe in some interesting situations. First, we analyze employment contracts based on education, training, or ability. We compare our results to those of the signaling literature (for example, Spence, 1973, 1974). Second, we consider optimal health insurance contracts and explain indemnity insurance. Finally, the problem of optimal compensation of law enforcers is discussed. Here we relate our results to those of Becker and Stigler.

A. *Ability, Training, and Employment Contracts*

There has recently been considerable interest in the relationship between education and other "signals" and the allocation of labor in job markets. For example, Spence (1973, 1974, 1976) stresses the use of education as a signal for native ability in the hiring of employees (see also Riley, Stiglitz, 1975b). In this literature, a particular payment structure for the worker, based on the signal, is assumed. This structure involves paying the worker a fixed wage which de-

pends only on his education level. These papers provide no analysis to justify the use of such a payment schedule, nor do they consider the possibility that the use of the signal in this way may be Pareto inferior to some other contractual arrangement. The present section is devoted to a clarification of this issue. We characterize the Pareto optimal contract in several situations similar to those analyzed in the signaling literature. In particular we exhibit conditions under which an *optimal* contract will depend on the education level of the worker (as well as things like recommendations, transcripts, previous experience, etc.), even when education is not a perfect measure of acquired ability or productivity. Thus our results may be viewed as complementary with the signaling literature.

To apply our model to the present problem, we reinterpret the effort as the ability the worker has acquired to perform the task, and the state as his native ability. We assume that neither the worker nor his prospective employer knows the worker's native ability at the time the contract is agreed on. Finally, each worker is assumed to know his own *acquired* ability.

With this interpretation, our results of Section II are applicable and imply:

i) If native ability is known or observable *ex post*, it follows from Propositions 1 and 2 that the optimal contract specifies that the employer receives an amount which depends only on the worker's native ability. In particular, the employer specifies a standard output to be produced by the worker contingent on his native ability. The employer's payoff is a certain share (which may depend on the worker's native ability) of this "standard output" while the worker receives the remainder of the standard revenue plus any output generated in excess of the standard (or minus any shortfall of actual output from standard output). This arrangement is equivalent to specifying a standard level of acquired ability required from a worker in this particular job and allowing workers with different levels of acquired ability to participate and accept the full consequences if their acquired ability is different from the one required.

ii) If workers are risk neutral, from Proposition 3, the Pareto optimal contract will specify a given payoff to employers independent of output, the worker's acquired ability, or his native ability level. Therefore, the worker receives the entire output minus some constant, that is, the worker purchases the right to use the production function for a given price.

iii) If workers are risk averse, contracts which depend only on the output are inefficient relative to contracts which depend on the native ability and/or acquired ability of the worker as well as the output. Pareto-superior results can be obtained if the acquired ability is observable and is included in the contract. From Proposition 4 it follows that Pareto optimal contracts when only output and acquired ability are observable are of the form "workers with acquired ability level of (at least) a^* will be paid $S(x)$, others need not apply," where $S(x)$ is some function of the output. For example, if years of education is a perfect correlate of acquired ability, this result implies that it is Pareto inferior not to make employment and/or salaries contingent on the education of the worker. Even if acquired ability is not observable (nor is something perfectly correlated with it) there may be imperfect measures which are sufficiently precise to make using them Pareto efficient. Examples of such monitors include years of education, interviews, transcripts, recommendations, etc. In this case we have shown sufficient conditions under which it is Pareto inferior not to include such monitors of acquired ability. The contract used is of the forcing type and specifies a minimum required level of the observable monitor of acquired ability. The above analysis provides an explanation for compensation schedules which are functions of education level. According to our analysis it is optimal for workers to receive *fixed wages* contingent on education level when employers are risk neutral. If employers are risk averse, a worker's compensation will

depend on his output if he meets the education requirement.

B. *Health Insurance Contracts*

It has for some time been recognized that when insurance payments depend on a decision of the insured as well as the state of nature, then an optimal allocation of resources and risk will not be achieved by a simple arrangement in which the insured pays a given price (premium) in return for various payments contingent on the state of nature. This problem arises because the insured has an incentive to "overspend" on insured expenses. It has been called "moral hazard" in the insurance literature (see, for example, Arrow, chs. 5, 8, 9; Pauly, 1968). One way to mitigate the inefficiency is to impose some of the cost of medical care on the insured. Zeckhauser illustrates the tradeoff between risk spreading and incentives. He suggests that insurance payments should depend on the degree of illness as well as the cost of the associated health care. Pauly (1971) refers to this type of insurance as "indemnity" insurance. The results of Section II can be applied to the case of health insurance to show under what conditions moral hazard causes inefficiencies and how appropriate contractual arrangements such as indemnity insurance can resolve this problem. Therefore, the results of Zeckhauser and Pauly (1971) can be obtained as special cases of our results.

To apply our results, we view the insurer as the principal and the insured as the agent. The random state θ is interpreted as the degree of illness. There are two possible interpretations of the agent's action, a. First, a may be the level of preventive care purchased by the insured or a general decision made by him *before* the state is realized as to the amount and quality of health care to be purchased when and if he becomes ill. The second interpretation of the agent's action is as a decision made *after* becoming ill on the amount and quality of health care to purchase. These two interpretations correspond to the assumptions of Models 1 and 2, respectively. The payoff x is interpreted as the amount spent on health care.[4]

Applying the results of Section II yields the following:

i) If the insurer can observe the degree of illness of the insured θ, then from Propositions 1 and 2 the optimal insurance contract specifies a given amount to be paid by the insurer for each possible degree of illness, that is, indemnity insurance. The amount paid by the insurer is thus independent of the choice of medical care taken by the insured. The insured under this contract can choose a level of medical care which costs more or less than the amount which the insurer agrees to pay. Moreover, since the insured is obviously risk averse, Propositions 1, 2, and 3 imply that indemnity insurance is *strictly* Pareto superior to insurance based only on the cost of the medical care. All of the above results hold whether the choice of medical care is taken before or after the occurrence of an illness.

ii) If the degree of illness cannot be observed by the insurer, but the insured's choice of medical care can be observed, then from Proposition 3 there are contracts which depend on both the cost of medical care and the insured's choice of medical care which are *strictly* Pareto superior to contracts which depend only on the cost. This result holds whether the choice of medical care is taken before or after the occurrence of illness. Furthermore, when the insured chooses his level of medical care *before* the occurrence of an illness, from Proposition 4, the Pareto optimal contract (when only the total cost and the insured's choice of care are observable) stipulates that the insurer pays some share of the cost *provided that the insured chooses some prespecified level of medical care.* Otherwise, the insurer pays nothing. Even when the insured's choice of medical care is not itself

[4]Note that these interpretations of a and x require opposite assumptions on the signs of the first derivatives of U and V than were made in Section I. It is easy to check that all the results continue to hold.

directly observable, it may be optimal to employ some indirect and imperfect measure of the level of care (for example, frequency of visits to a doctor). When these measures are sufficiently precise estimates of the insured's actual choice of medical care, the optimal policy will be essentially the same as when the choice is directly observable.

C. *Compensation of Law Enforcers*

In a very interesting paper, Becker and Stigler analyze the law enforcement problem and suggest two alternative methods for improving the incentives given enforcers. Here we recast the law enforcement problem in our framework and employ the results of the previous section to obtain the optimal contract between the state and enforcers. Our results provide a firm foundation for the suggestions of Becker and Stigler. In particular, we show explicitly under what conditions each suggestion is Pareto optimal.

Becker and Stigler explore the economic incentives for malfeasance in law enforcement. They conclude that officials responsible for enforcement might lack sufficient incentives to enforce certain laws. Moreover, there may also be incentives for them to engage in malfeasance. Becker and Stigler suggest two methods for improving the incentives given enforcers. The first suggestion discourages malfeasance and lack of proper enforcement by penalizing the enforcer if such behavior is detected. The penalty is set such that it more than offsets the gain from malfeasance. This method is made operational by requiring the enforcers to "post a bond equal to the temptation of malfeasance, receive the income on the bond as long as they are employed, and have the bond returned if they behave themselves until retirement" (Becker and Stigler, p. 9). If the state detects malfeasance on the part of the enforcer, he is fired and loses the bond he posted. The second suggestion is to allow free entry into the enforcement industry. Enforcers would be rewarded based on their performance. Becker and Stigler argue that the amount of enforcement would be optimal if successful enforcers were paid the fines levied against convicted violators. These fines would equal the damages to society caused by the violator divided by the probability of conviction.

The situation described by Becker and Stigler can be stated in terms of Models 1 and 2 of Section I. The state (society) and the enforcer can be viewed as the principal and agent, respectively. The payoff x to the law enforcement activity is the revenue generated via the fine levied on convicted violators net of the costs associated with trying the accused. These latter costs include both costs borne by the accused (for example, time lost) and those borne by the state (for example, court costs). As is the case in the Becker-Stigler paper, we are not concerned with the problem of *crime prevention* or the effects of law enforcement on the level of criminal activity. In fact, we assume here that the level and type of criminal activity is completely exogenous. The schedule of fines is also taken to be exogenous. The state is interested in the degree of law enforcement as measured by the revenue generated from fine collection.

Two important aspects of the incentive problem for law enforcers may be distinguished. These aspects correspond to two distinct interpretations of the agent's action a, and the exogenous random variable θ (recall that a and θ jointly determine the payoff x). In the first, the effort of the enforcer can be interpreted as his level of investment in crime detection and apprehension capabilities and the levels of activities such as patrolling, etc. These decisions must occur *before θ, the level of crime activity, is known.* In this case, the problem is to provide, in an efficient contract, the proper incentives for investment in crime detection and apprehension capabilities. This problem may be analyzed using Model 1, since the agent's action (investment and patrolling) is taken before the realization of the exogenous random state (level of criminal activity).

In the second interpretation, the random state θ is the type of crimes committed, the

identities of the criminals, and other details associated with the crimes (for example, evidence, etc.). The action of the enforcer a is the effort expended in apprehending the criminals, creating cases, and the extent to which the enforcer refrains from engaging in malfeasance (for example, taking bribes, etc.). We assume for the purpose of this analysis that the enforcer knows all the details of the crimes (including the identities of the criminals) before taking his action. Society, as represented by the state, may or may not know these details. In any case, we assume contracts are agreed upon before crimes are committed. This version of the problem can be analyzed using Model 2.

Our results indicate the following:

i) With regard to the first problem, suppose the level of criminal activity can be observed *ex post*, both by the state and by enforcers. In this case, it follows from Propositions 1 and 2 that the optimal contract specifies that the state receives an amount which depends only on the level of criminal activity. In particular, the state specifies a standard for the revenue generated by enforcement activities contingent on the level of criminal activity. It then receives a certain share of this standard revenue while enforcers receive the remainder of the standard revenue plus any revenues generated in excess of the standard (or minus any shortfall of actual revenue from standard revenue). Essentially, the optimal arrangement is equivalent to one in which the state specifies a given level of investment in enforcement capability and enforcers accept full responsibility for any deviations from this level.

Regarding the second problem, when both the state and enforcers can observe the particulars of the crimes which occur, Propositions 1 and 2 imply essentially the same result but with a slightly different interpretation. Here the state specifies a total amount to be recovered by the enforcer, contingent on the particular circumstances of the crimes which have occurred. The state receives a share of this specified amount, independent of the enforcer's action. The enforcer receives the remainder of the specified amount, plus or minus any deviation of the actual amount recovered from the specified amount. All consequences of malfeasance are borne by the enforcer.

ii) If law enforcers are risk neutral, as assumed by Becker and Stigler, the Pareto optimal contract in both interpretations specifies that enforcers receive the entire payoff minus, perhaps, some constant. This follows directly from Proposition 3 and is similar to Becker and Stigler's second suggestion. In this case, enforcers purchase for a fixed amount the right to enforce laws and collect fines. There is no point in observing either the enforcer's action or the random state when the enforcers are risk neutral.

iii) If law enforcers are risk averse, contracts which depend only on the revenue generated may be inefficient. In the second interpretation, by Proposition 3, Pareto-superior results can be achieved if the particulars of the crimes committed can be observed by both parties. In this case the particular form of the Pareto optimal contract is as described in (i). Also, if there is a one-to-one relationship between the payoff and the crimes committed for any given action by the enforcer, then Pareto-superior results can be obtained if the enforcer's action can be observed.

In the first interpretation, Pareto-superior results can be achieved if either the action or the level of criminal activity can be observed by both parties. Moreover there are potential gains to monitoring (imperfectly) the activities of enforcers including detection of shirking. Under certain conditions on the monitoring technology, forcing contracts will be Pareto optimal. Under this type of contract, the enforcer would receive an amount which depends on the payoff if monitoring reveals that his action is "acceptable" (for example, there was no shirking). If his action is found to be unacceptable, he would receive a smaller amount or pay a penalty. This contract is similar to the first method suggested by Becker and Stigler, in which enforcers are required to post a bond. In order for such an arrangement to be Pareto superior to compensation

which depends only on the revenue, the state must possess means of detecting unacceptable performance by enforcers which have low probabilities of error. The restrictions on the quality of the detection mechanism needed to guarantee Pareto superiority of the Becker-Stigler proposal are quite severe.

REFERENCES

A. Alchian and H. Demsetz, "Production, Information Costs, and Economic Organization," *Amer. Econ. Rev.*, Dec. 1972, *62*, 777–95.

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing,* Chicago 1970.

G. S. Becker and G. J. Stigler, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *J. Legal Stud.*, Jan. 1974, *3*, 1–18.

M. Harris and A. Raviv, "Optimal Incentive Contracts with Imperfect Information," Grad. Sch. Ind. Adm., work. paper no. 70-75-76, Carnegie-Mellon Univ., April 1976.

C. M. Jensen and W. H. Meckling, "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *J. Finan. Econ.*, Oct. 1976, *3*, 305–60.

M. V. Pauly, "The Economics of Moral Hazard," *Amer. Econ. Rev.*, June 1968, *58*, 531–37.

———, "Idemnity Insurance for Health Care Efficiency," *Econ. Bus. Bull.*, Fall 1971, *24*, 53–59.

J. G. Riley, "Competitive Signalling," *J. Econ. Theory*, Apr. 1975, *10*, 174–86.

S. A. Ross, "The Economic Theory of Agency: The Principal's Problem," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 134–39.

A. Michael Spence, *Market Signaling: Information Transfer in Hiring and Related Processes*, Cambridge, Mass. 1973.

———, "Competitive and Optimal Responses to Signals: An Analysis of Efficiency and Distribution," *J. Econ. Theory*, Mar. 1974, *8*, 1296–332.

———, "Competition in Salaries, Credentials, and Signaling Prerequisites for Jobs," *Quart. J. Econ.*, Feb. 1976, *90*, 51–74.

——— **and R. Zeckhauser,** "Insurance, Information, and Individual Action," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 380–87.

J. E. Stiglitz, (1975a) "Incentives, Risk and Information: Notes Toward a Theory of Hierarchy," *Bell J. Econ.*, Autumn 1975, *6*, 552–79.

———, (1975b) "The Theory of 'Screening', Education, and Distribution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

R. M. Townsend, "Efficient Contracts with Costly State Verification," Grad. Sch. Ind. Admin., work. paper no. 14-77-78, Carnegie-Mellon Univ. 1976.

R. B. Wilson, "On the Theory of Syndicates," *Econometrica*, Jan. 1968, *36*, 119–32.

———, "The Structure of Incentives for Decentralization Under Uncertainty," in *La Decision*, Paris 1969.

R. Zeckhauser, "Medical Insurance: A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives," *J. Econ. Theory*, Mar. 1970, *2*, 10–26.

Reliability and Public Utility Pricing

By M. A. CREW AND P. R. KLEINDORFER*

The problem of public utility pricing under risk has been the subject of several articles in this *Review*. Gardner Brown, Jr. and M. Bruce Johnson (B-J) set the stage by showing, *inter alia*, that peak load pricing might be nonoptimal for a welfare-maximizing monopolist. As the B-J results were at odds with Peter Steiner's long-accepted results for the deterministic case, a lively discussion ensued.[1] One feature of this discussion was the recognition that the B-J framework had neglected the important issue of reliability and that, perhaps, imposing explicit constraints limiting the probability of excess demand might significantly alter their results. This issue lay dormant until Robert Meyer (1975) provided a normative framework for monopoly pricing under uncertainty. Meyer's conclusions confirmed the earlier suspicions that B-J's results could be seriously off the mark if reliability constraints were imposed. However, Meyer left open the key question of how such reliability levels should be chosen, either for a welfare-maximizing or for a regulated profit-maximizing monopolist. The purpose of this paper accordingly is to provide some rationale for the choice of such reliability levels.

Section I describes the basic problem statement and first-order conditions. The pricing results of B-J for a welfare-maximizing monopolist are reexamined as well as some considerations affecting Meyer's extension of these to the profit-maximizing case. We show that there are in general multiple optima to the welfare-maximizing problem, and that the B-J solution is the particular optimal solution which minimizes net revenue. This nonuniqueness arises from the fact that the familiar welfare criterion of maximizing the sum of consumer's and producer's surplus provides no basis for choosing among alternative distributions of welfare between consumers and producers. This is changed when reliability constraints are added.

Section II analyzes reliability constraints. We provide a framework for relating the optimal choice of such reliability levels to the costs of excess demand. Section III examines some numerical examples illustrating the relationships between risk, pricing policies, capacity levels, and reliability levels for both welfare-maximizing and profit-maximizing monopolists. Finally, Section IV provides a summary and a discussion of the related issue of regulation.

I. Problem Formulation and First-Order Conditions

The problem of concern is that of public utility pricing under uncertainty.[2] After restating the known first-order conditions for this problem, we turn our attention to an apparent discrepancy between previous results obtained for the stochastic and deterministic cases. In the former, the B-J welfare-maximizing solution is $p = b$, while the familiar deterministic solution is $p = b + \beta$, where β is the constant cost of supplying a unit of capacity and b is (constant) marginal running cost per unit per period. Apparently, then, the presence of uncertainty, no matter how small, leads to a radically different pricing rule than that appropriate when uncertainty is not present. We

*Associate professor of business administration, Rutgers University, and professor of decision sciences, University of Pennsylvania, respectively. We are grateful to seminar participants at Bonn University, Brown University, Erlangen-Nurnberg University, Oklahoma State University, and Bell Laboratories for illuminating comments on an earlier draft of this paper. We would also like to thank George Borts, Howard Kunreuther, Robert Meyer, Bridger Mitchell, Roger Sherman, and a referee for helpful advice and comments.

[1]See especially Ralph Turvey.

[2]See Brown and Johnson, Meyer (1975), and the authors (1976).

resolve this issue below by showing that there are generally multiple optima to the pricing problems considered, and that B-J and Steiner simply considered different (and incompatible) optima to their respective problems. As it turns out, the *set* of optimal solutions to the B-J problem converges as uncertainty decreases to zero to the *set* of optimal solutions of the corresponding (Steiner) problem under certainty. Thus, where uncertainty is small, deterministic problems do in fact serve as reasonable approximations for corresponding problems under uncertainty.

Restricting attention to the 2-period case, we denote demand for each of two equal length periods as $D_i(p_i) + \tilde{u}_i$, where $\tilde{u}_i$ is a random variable with mean zero and where the mean demand function D_i is negative sloping with inverse D_i^{-1}. Let Q be the amount of capacity installed with capacity and running costs as above. Representing actual output in period i by S_i, we see that S_i is the minimum of demand and capacity, that is, for any given value of the disturbance term[3] $\tilde{u}$, say $\tilde{u} = (u_1, u_2)$, and for any price vector $p = (p_1, p_2)$ and capacity Q, we have

$$(1) \quad S_i(p_i, Q, u_i) = \mathrm{Min}[D_i(p_i) + u_i, Q]$$

The welfare returns obtained for given u, p, and Q are therefore[4]

$$(2) \quad W(p, Q, u) = \sum_{i=1}^{2} (p_i - b) S_i(p_i, Q, u_i) - \beta Q + \sum_{i=1}^{2} \left(\int_0^{S_i(p_i, Q, u_i)} [D_i^{-1}(x - u_i) - p_i] \, dx \right)$$

with corresponding returns $\pi(p, Q, u)$ for a profit maximizer being the same except for the omission of the third term in (2).

Denoting the expected value operator by E, our interest then is to maximize $EW(p, Q, \tilde{u})$ (or $E\{\pi(p, Q, \tilde{u})\}$) over the set of nonnegative price vectors p and capacities Q. First-order conditions for these problems have been derived under various conditions by several authors.[5] For the case of a welfare maximizer, these turn out to be the following when $\tilde{u}_1$ and $\tilde{u}_2$ are independent:

$$(3) \quad (p_i - b) G_i[Q - D_i(p_i)] = 0, \qquad i = 1, 2$$

$$(4) \quad \sum_{i=1}^{2} \left(\int_{Q - D_i(p_i)}^{\infty} [D_i^{-1}(Q - u_i) - b] \cdot g_i(u_i) \, du_i \right) = \beta$$

where g_i is the density function and G_i is the cumulative distribution function (CDF) of the random variable u_i. We note in passing that, since $G_i(u) = Pr\{\tilde{u}_i \leq u\}$ by definition of a CDF, $G_i[Q - D_i(p_i)] = Pr\{\tilde{u}_i \leq Q - D_i(p_i)\} = Pr\{D_i(p_i) + \tilde{u}_i \leq Q\}$ is just the reliability in period i.

Now the B-J results are obtained by canceling the term $G_i[Q - D_i(p_i)]$ in (3), giving the solution $p_i = b$ for all periods. As we indicate directly below, however, reliability may be zero so that canceling $G_i[Q - D_i(p_i)]$ which is the reliability in period i is not always justified. This is the source of the difficulties noted above concerning the B-J results.

Suppose the policy (p, Q) is chosen. Then when $\tilde{u}_i = u_i$, the (marginal) loss in consumer surplus and revenue over variable cost in period i from not increasing Q slightly is given by $(D_i^{-1}(Q - u_i) - b)$ if capacity is exceeded (i.e., if $D_i(p_i) + u_i \geq Q$) and is zero otherwise. Thus, equation (4) can be interpreted as: set Q just large enough so that the marginal expected losses due to unmet demand are equal to marginal capacity cost β. To accomplish this when uncertainty is low or β is large would in

[3]Our notational convention is to place a tilde over random variables.

[4]Note that this presumes the B-J rationing policy which assumes that, in the event of demand exceeding capacity, it is possible to rank consumers according to their marginal willingness to pay. This is necessary from a technical point of view for the values of the integral to reflect consumer's surplus, but it has no value-free justification. It does make the B-J analysis comparable with the assumption of the deterministic analysis that income distribution questions are not to be considered in this type of optimal pricing analysis, as clearly explained in Oliver Williamson.

[5]See Brown and Johnson, Meyer (1975), and the authors (1976).

general require either capacity to be small or one of the prices (obviously the peak price) to be higher than b, or both. If one continues to insist on $p_i = b$ as a pricing policy, then the marginal argument embodied in (4) will compel Q to become so small as uncertainty decreases that eventually reliability will become zero; i.e., $G_i[Q - D_i(b)] = Pr\{\text{demand does not exceed } Q \text{ in period } i\}$ will equal zero.[6] At this point the cancellation of this term in (3) is no longer permissible, and thus the result that $p_i = b$ would not necessarily follow for low levels of uncertainty and/or high capacity costs.

The above intuitive line of reasoning can be strengthened by a short technical argument. To do so, we assume that period 2 is a firm peak period in the usual deterministic sense of Steiner. We then look for optimal solutions satisfying $p_1 = b$ and

$$\text{(5)} \qquad \text{Reliability in Period 1} = G_1[Q - D_1(p_1)] = 1$$

$$\text{(6)} \qquad \text{Reliability in Period 2} = G_2[Q - D_2(p_2)] = 0$$

Equation (5) implies absolute reliability in the off-peak period, the certainty of meeting demand; equation (6) implies absolute unreliability, the certainty that there will be some excess demand in the peak period. If (5) and (6) can be satisfied, then the lower limit of integration for period 1 (respectively, 2) in (4) becomes effectively $+\infty$ (respectively, $-\infty$) and (4) reduces to[7]

$$\text{(7)} \qquad E\{D_2^{-1}(Q - \tilde{u}_2)\} = b + \beta$$

a condition on Q alone. Thus, if (5)–(7) can be satisfied with $p_1 = b$, it is clear that (3)–(4) hold and the resulting solution is optimal.[8]

As uncertainty decreases to zero, the expected value in (7) approaches[9] $D_2^{-1}(Q)$, that is, (7) becomes $D_2^{-1}(Q) = b + \beta$, the solution of which is the well-known deterministic optimal capacity, $Q_d = D_2(b + \beta)$. Moreover, since period 2 is a firm peak and $D_2(b + \beta) = Q_d$, we have $D_1(b) < Q_d < D_2(b)$, so that for uncertainty sufficiently small the B-J solution, $p_i = b$ for $i = 1, 2$, will satisfy (5)–(6).

Note, however, that for low levels of uncertainty the peak price can be increased above b (until roughly[10] $G_2[Q_d - D_2(p_2)]$ first becomes positive) without violating (5)–(7). Thus, a range of solutions obtains for the optimal peak price. To obtain all such optimal solutions, one simply solves (7) for the optimal capacity Q. Then, setting $p_1 = b$, one insures that (5) is satisfied, which will be the case for uncertainty sufficiently small as we have just pointed out. Then solve for all $p_2 \geq b$ satisfying (6). If no such p_2 exists, then $p_2 = b$ is the only optimal solution (as indicated by (3)). Otherwise any such p_2 is optimal.

Since satisfying (6) is easiest when $D_2(b) \gg Q$, the range of peak prices satisfying (6) decreases as uncertainty increases and/or as the excess of peak demand at the price b over deterministic optimal capacity decreases (note that $Q \cong Q_d$ for low levels

[6]We are assuming here that $\tilde{u}_i$ is bounded from below, otherwise $G_i[Q - D_i(p_i)]$ would always be positive and the B-J result would always apply. However, an unbounded distribution could have no sense since it would imply negative quantities demanded. We will assume henceforth, with assuredly no loss in practical generality, that the disturbance terms $\tilde{u}_i$ are bounded both above and below.

[7]To derive (7) simply note that when (5) is satisfied $\tilde{u}_1$ assumes values no greater than $Q - D_1(p_1)$. Thus, the corresponding period 1 integral in (4) is zero. Similarly (6) implies that $\tilde{u}_2$ is always no less than $Q - D_2(p_2)$, so that the period 2 integral in (4) is just $E[D_2^{-1}(Q - \tilde{u}_2) - b]$. Together these facts yield (7).

[8]As noted in the authors (1976), conditions (3)–(4) are also sufficient for optimality for the expected welfare-maximizing problem, so the search for such optima may be restricted to solutions fulfilling (3)–(4).

[9]See Patrick Billingsley, p. 24, for a proof of this, assuming that $\tilde{u} \rightarrow 0$ means convergence in distribution to a random variable having a degenerate distribution centered at the origin.

[10]The reader should note that when demands are linear the left-hand side of (7) is exactly $D_2^{-1}(Q)$, since $E(\tilde{u}_2) = 0$. Thus, when demands are linear and uncertainty is sufficiently low there is a range of levels of uncertainty over which the optimal capacity decision is precisely Q_d. When demands are not linear, optimal capacity may be greater than or less than deterministic optimal capacity depending on the concavity or convexity of the demand curve. That is, $E\{D_2^{-1}(Q - \tilde{u}_2)\}$ may be less than or greater than $D_2^{-1}(Q - E(\tilde{u}_2)) = D_2^{-1}(Q)$.

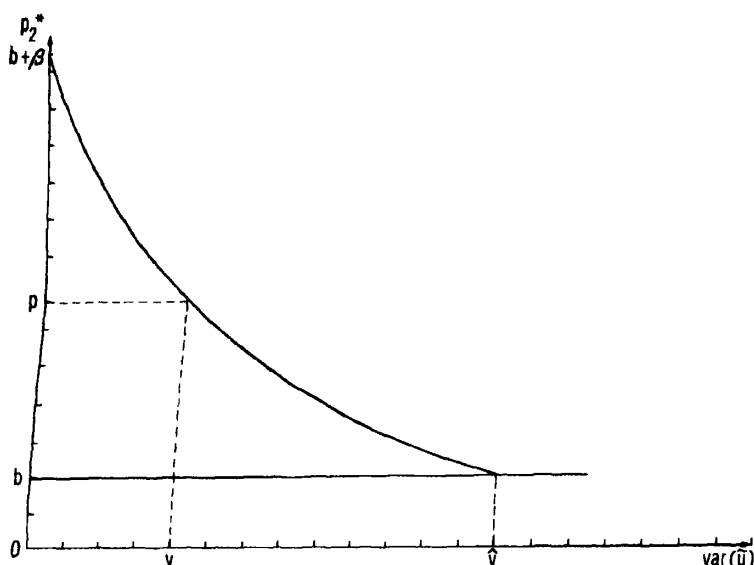



FIGURE 1. OPTIMAL PEAK PRICE (p_2^*) AND VAR ($\tilde{u}$)

of uncertainty). In particular, the range of possible optimal peak prices increases as β increases.

Figure 1 illustrates the outcome of the above discussion. There we plot optimal peak price versus the variance of u_t (the optimal off-peak price is, of course, b). Up until $\hat{v}$, it is seen that a (decreasing) range of optimal peak prices obtains. Thereafter, only the B-J solution is optimal. Note that when there are multiple optima, each such optimal solution yields the same expected welfare, but differential revenue effects. Considering the point v in Figure 1, for example, revenue and profits increase in going from the B-J solution to the maximal optimal solution ($p_1 = b$ and $p_2 = p$), while expected consumer's surplus decreases correspondingly.

Considering the deterministic case, it is interesting to note that $p_1 = p_2 = b$ is indicated to be optimal. The fact that this solution has not generally been recognized as optimal stems from the fact that previous derivations for the deterministic case have only considered solutions satisfying the constraint that sufficient capacity be installed to meet peak demand.[11] In the absence of such a (reliability) constraint, it may easily be verified that the B-J solution is indeed "optimal" for the familiar welfare criterion which does not discriminate between producer's and consumer's surplus.[12] It is in fact that optimal solution which provides minimum profit, minimum revenue, and maximum consumer's surplus and excess demand.

Returning to the stochastic case, the B-J solution continues to enjoy all of the above properties. It is indeed optimal, but among the possible optima, it is surely not what common sense would dictate as the optimum optimorum, particularly as relates to reliability. It therefore seems reasonable to introduce some form of reliability constraint into the problem, as Turvey and Meyer (1975) have suggested, in order to avoid the obvious difficulties of a pricing policy which could induce very low reliability and net revenues. In order to do so in a rational manner, we introduce the notion of

[11]Robert Dansby has made a similar point in his analysis of pricing decisions for time-varying demand.

[12]In fact, the reader may verify using the usual Steiner analysis that the absence of a reliability constraint leads to $p_1 = b$ and *any* peak price between 0 and $b + \beta$ being optimal. Thus, pricing reversals are possible in the welfare-maximizing case provided that one is willing to have zero reliability in the peak period. We disregard in the sequel optimal peak prices which do not at least recover operating expenses.

rationing costs and relate these to optimal reliability levels.

II. Rationing Costs and Reliability Constraints

As noted earlier, for the integration in (2) to measure consumer's surplus accurately, it is necessary to assume that in the event of demand exceeding capacity, consumers will be ranked according to marginal willingness to pay. We employ this assumption, but unlike B-J we assume that this process is not costless. In the event of scarcity, as Roger Sherman and Michael Visscher, and Visscher have also indicated, costs arise in ensuring that consumers are rationed in accordance with their marginal willingness to pay. For example, for a utility employing a two-part tariff[13] some negotiating process might take place between the public utility and its consumers whereby it agrees to "pay" those consumers who will be cut off in the event of demand exceeding capacity. While not employing such a scheme, we contend that there are costs in rationing according to willingness to pay in the event of excess demand. Moreover, as the quantity of excess demand increases it is likely that the transactions costs just from dealing with more consumers will increase. We accordingly define the sum of such costs when capacity is exceeded as

$$(8) \qquad r = \sum_{i=1}^{2} r_i(D_i(p_i) + \tilde{u}_i - Q)$$

We assume that r_i is a monotonic increasing, convex, and differentiable function of excess demand.

A similar argument to the above can be made for the existence of rationing costs for the profit-maximizing case. Although the profit maximizer is not necessarily concerned with willingness to pay as a rationing criterion, he is nonetheless confronted with the task of organizing and administering a rationing scheme in the event of excess demand, and the costs of this task may be assumed to have the general properties of r_i just given. A further likely cost of excess demand in a regulated environment is the cost of probable audits as well as other longer-run negative profit effects of chronic excess demand.[14] We will assume such costs have been incorporated in r_i when dealing with the profit-maximizing case. Our principal interest in this section will be, however, the welfare-maximizing case.

We subtract (8) from (2) to obtain a new measure of welfare returns:

$$(9) \qquad W'(p,Q,u) = W(p,Q,u) - \sum_{i=1}^{2} r_i(D_i(p_i) + u_i - Q)$$

Following the procedure of Section I we maximize the expected value of (9), which we denote by $\overline{W}'$. By differentiating[15] this expected value with respect to p_i we have

$$(10) \qquad \frac{\partial \overline{W}'}{\partial p_i} = (p_i - b)G_i[Q - D_i(p_i)] - E\{r_i'(D_i(p_i) + \tilde{u}_i - Q)\}$$

From (10) if $G_i[Q - D_i(p_i)] = 0$, then demand always exceeds capacity implying that $E\{r_i'(D_i(p_i) + \tilde{u}_i - Q)\} > 0$, which violates the optimality condition that (except in the uninteresting case of $p_i = 0$), $\partial \overline{W}'/\partial p_i = 0$. Thus, the presence of rationing costs always

[13]It is interesting to note in this regard that the claim by Meyer (1975) that his framework accomodates multiple groups of users making demands on a common facility at a single point in time is evidently false. For this to be so, his capacity constraints would have to read $\Sigma D_i(p_i) \leq Q$, whereas his are always of the form $D_i(p_i) \leq Q$ (no summation). Clearly, the relevant literature here is that relating to multipart tariffs, for example, Elizabeth Bailey and Lawrence White, and Walter Oi.

[14]For a discussion of the threats posed by such audits or "reviews" see V.S. Bawa and David Sibley. These rationing costs and longer-run effects are presumably what Meyer (1975) had in mind in suggesting that a profit-maximizing firm might be risk averse to excess demand. Nonetheless, his exposition of this issue is somewhat unsatisfying because it mixes into the decision maker's utility function money and excess demand. The decision maker is apparently risk neutral in the former while he is risk averse in the latter. The approach suggested here is to account for all cost, revenue, and rationing effects monetarily, and then to express whatever risk aversion might be present in terms of a utility function on the total (stochastic) monetary outcome of a particular policy. For reasons of space we will be dealing solely with the risk-neutral case, however.

[15]See the authors (1976) for a derivation of (10).

insures that $G_i[Q - D_i(p_i)]$ is positive in contrast to the B-J solution. Note also that in case of any excess demand, expected marginal rationing costs are positive and (10) implies $p_i > b$. Finally, as rationing costs are likely to be higher in the peak period, price would also be higher there, implying peak load pricing when rationing costs are present.[16]

Results similar to these have been derived by Meyer (1975) who suggests maximizing the traditional welfare function subject to reliability constraints, that is,

$$(11) \quad \text{Max}\, \overline{W}(p, Q, \tilde{u}) = E\{W(p, Q, \tilde{u})\}$$

subject to

$$(12) \quad G_i[Q - D_i(p_i)] \geq \epsilon_i, \qquad i = 1, 2$$

where $\epsilon_i (0 \leq \epsilon_i \leq 1)$ are the specified reliability levels and where the welfare function W in (11) is defined in (2).

Now suppose the reliability levels ϵ_i in (12) are set to $\epsilon_i = \hat{\epsilon}_i = G_i[\hat{Q} - D_i(\hat{p}_i)]$, with $(\hat{p}, \hat{Q})$ an optimal solution to the rationing cost problem of maximizing the expected value $\overline{W}' = \overline{W} - \Sigma E\{r_i\}$ of (9). Since rationing costs are monotonic increasing and convex, it is clear that the higher the required reliability levels ϵ_i in (12) the lower will be the expected rationing costs at optimum. Therefore, (12) may be regarded as a constraint on expected rationing costs and if the reliability levels ϵ_i are (optimally) specified as $\epsilon_i = \hat{\epsilon}_i$, then the solution $(\hat{p}, \hat{Q})$ to maximizing W also solves the reliability constrained problem (11)-(12).

Thus, reliability constraints may be thought of as a surrogate for rationing costs, provided of course that they are specified optimally. We will return to this point below in discussing regulation. For the moment, though, we may note that either rationing costs or nonzero reliability constraints will preclude the behavior at optimum exhibited by the B-J solution. To illustrate this we turn to some numerical examples.

[16]See the authors (1976) for a more precise analysis of the optimality of peak load pricing under uncertainty.

III. Some Illustrative Examples

Our basic model for these examples is for two periods. The following demand functions are assumed:

$$(13) \quad D_1(p_1, \tilde{u}_1) = 40 - \frac{1}{7} p_1 + \tilde{u}_1$$

$$(14) \quad D_2(p_2, \tilde{u}_2) = 80 - \frac{5}{7} p_2 + \tilde{u}_2$$

We assume that the disturbance terms $\tilde{u}_i$ are uniformly distributed on $[-\gamma, \gamma]$, where γ is a positive constant identical in both periods. We assume that plant costs are $b = 5$, $\beta = 10$. The results given in Table 1 are confined to the case where the effects of rationing costs are represented through reliability constraints.[17] The same (minimum) reliability level (ϵ) was required for both periods.

Let us first consider the effects of increasing uncertainty on the level of capacity and price where no reliability constraint is imposed (the B-J case in the welfare-maximizing case). We see that as the range of the disturbance term (γ) increases, for both the profit-maximizing and the welfare-maximizing objective functions, optimal capacity increases; the profit-maximizing capacity, however, always being less than the welfare-maximizing capacity. The levels of maximum profit implied at each level of γ vary little (decreasing from 4381 to 4215 as γ increases from 0 to 20), in contrast with the optimal welfare solution, where profits vary between 0 and -725. The welfare-optimal pricing policies feature multiple optima for the peak price for a wide range of values of the disturbance term as previously indicated in Figure 1. When the variance of the disturbance is zero, the deterministic case,

[17]A direct search method was used in solving these examples. Details of the solution procedure as well as further numerical results will appear in the authors' forthcoming book. For some illustrative numerical results on the case where rationing costs are positive, see the authors (1976). As expected from the discussion in Section II, these turn out to be entirely symmetric to those presented here using reliability constraints, where higher rationing costs correspond to higher (optimal) reliability levels.

TABLE 1—OPTIMAL SOLUTIONS FOR THE EXAMPLE

Reliability Level Required		Range of Disturbance Term (γ)					
		Welfare-Maximizing Case			Profit-Maximizing Case		
		$\gamma = 0$[a]	$\gamma = 5$	$\gamma = 15$	$\gamma = 0$[a]	$\gamma = 5$	$\gamma = 15$
$\epsilon = 0$[b]	p_1	5.00	5.00	5.00	142.40	142.50	142.50
	p_2	5.-15.0	5.-8.0	5.00	63.50	63.39	63.19
	Q	69.29	69.29	70.73	34.65	38.00	44.71
$\epsilon = .5$	p_1	5.00	5.00	5.00	142.50	142.50	142.50
	p_2	15.00	13.24	9.74	63.50	63.39	63.19
	Q	69.29	70.54	73.04	34.65	38.00	44.71
$\epsilon = .9$	p_1	5.00	5.00	5.00	142.50	142.50	142.50
	p_2	15.00	14.92	14.77	63.50	63.45	63.40
	Q	69.29	73.34	81.44	34.65	38.67	46.76
$\epsilon = .98$	p_1	5.00	5.00	5.00	142.50	142.50	142.50
	p_2	15.00	14.99	14.98	63.50	63.49	63.49
	Q	69.29	74.09	83.70	34.65	39.45	49.05
$\epsilon = 1.0$	p_1	5.00	5.00	5.00	142.50	142.50	142.50
	p_2	15.00	15 00	15.00	63.50	63.50	63.50
	Q	69.29	74.29	84.29	34.65	39.65	49.65

[a]The Deterministic Case
[b]The B-J Case

any peak price between 5 and 15 is optimal. The range of optimal peak prices decreases as γ increases until at just over $\gamma = 7$ it converges to 0, with the only optimum being the B-J solution of $p_1 = p_2 = b = 5$.

Now consider the effects of increasing reliability requirements on optimal prices and capacities. As reliability (ϵ) is increased, capacity stays constant for a while (until the optimum unconstrained reliability level is reached) and then begins to increase, the sharpness of the increase depending on the level of uncertainty, with the welfare case requiring more capacity than the profit case. For the optimal prices it is interesting to note that they approach the deterministic optimal prices as the required reliability approaches 100 percent. This follows because very high reliability levels require the utility to plan "deterministically" on meeting (nearly) all demand. The optimal prices for the profit-maximizing case are less interesting. The optimal peak period price varies only between $63.19 \leq p_2 \leq 63.50$ and $63.39 \leq p_2 \leq 63.50$ for the cases $\gamma = 15$ and $\gamma = 5$, respectively, with the off-peak price $p_1 = 142.5$ in both cases. Perhaps the only interesting point about this is that it illustrates that peak/off-peak reversals can occur under stochastic demand. Nor is the level of profit affected much by changes in the required reliability level. For reliability levels between zero and one it varies by a fraction of 1 percent, implying for this case that, aside from considerations of variance of profit,[18] a profit maximizer will not be much concerned about reliability unless there exists some means of insuring that he bears some of the social costs of failure to meet demand.

IV. Summary and Conclusions

We may now summarize the main thrust of the argument. In deterministic models of public utility pricing it has been traditional to impose the constraint that peak demand should exactly equal capacity. This may be viewed as a reliability constraint. It may also be viewed as a minimum profit or revenue constraint. If such a constraint is not imposed, multiple pricing optima occur due

[18]Meyer (1976), by drawing on the capital assest pricing literature, has suggested one possible way of incorporating such considerations.

to the nature of the welfare function. Precisely the same properties hold for the stochastic optimal solution when uncertainty is sufficiently low, with all resulting multiple optima entailing zero reliability. Furthermore, some numerical examples presented indicate that such effects may persist over a very wide range of levels of uncertainty. Since common sense implies a fairly high reliability level, this suggests that some appropriate representation of the traditional deterministic reliability constraint (that $D_2(p_2) = Q$) is an important aspect of public utility pricing under uncertainty.

To provide some rationale for such reliability constraints we have introduced explicitly the costs of rationing when excess demand occurs. This was shown to be formally equivalent to selecting optimal reliability levels. The presence of such rationing costs then induces optimal behavior close to that implied by the usual deterministic analysis. In particular, peak load pricing continues to be optimal in the stochastic case since marginal rationing costs[19] are higher in peak periods. Another interpretation of why the stochastic solution is similar to the deterministic solution is that the presence of rationing costs implies a high optimal reliability level and results in the utility being confronted with the essentially deterministic problem of meeting a very high fractile of the demand distribution.

The above results also have certain implications for the regulatory environment. Again here we note that deterministic models[20] of regulation of profit-maximizing monopolies typically contain the same reliability constraint as in the welfare-maximizing case—namely, that capacity be precisely equal to peak demand. The corresponding representation of this constraint under uncertainty could again be some probabilistic reliability constraint which, if sufficiently close to one, would doubtless induce deterministic-like behavior in a risk-neutral environment. However, for the regulated firm, it seems even more crucial to have an explicit rationale for analyzing such reliability constraints, since these are likely to interact strongly with other regulatory constraints. Although a full analysis of this issue is beyond the scope of this paper, a simple example should indicate the dangers in a regulatory environment of neglecting reliability or treating it in an *ad hoc* manner.

Consider the following model[21] of an expected rate-of-return regulated monopoly:

$$\text{(15)} \qquad \text{Max}\, E\{\pi(p, Q, \tilde{u})\}$$

subject to

$$\text{(16)} \qquad E\{\pi(p, Q, \tilde{u})\} \leq (s - \beta)Q$$

where $p \geq 0$, $Q \geq 0$, $s - \beta > 0$, where s is the maximum rate of return allowed by the regulatory commission, and where $E\{\pi\}$ is defined through

$$\text{(17)} \quad E\{\pi\} = E\left\{\sum_{i=1}^{2} (p_i - b)S_i(p_i, Q, \tilde{u}_i) + r_i(D_i(p_i) + \tilde{u}_i - Q)\right\} - \beta Q$$

Now suppose that the disturbance term $\tilde{u}$ is bounded, that is, there is some bounded interval $[u_{1i}, u_{2i}]$ such that $Pr\{\tilde{u}_i \in [u_{1i}, u_{2i}]\} = 1$, and let the maximum expected revenue obtainable when capacity is effectively a free good be denoted by R^*. When capacity Q is a free good, the monopolist will set Q high enough to insure that demand is always less than Q. Thus,

$$\text{(18)} \quad R^* = \underset{p \geq 0}{\text{Max}}\, E\left\{\sum_{i=1}^{2} (p_i - b) \cdot [D_i(p_i) + \tilde{u}_i]\right\}$$

or, since $E\{\tilde{u}_i\} = 0$, R^* reduces to the maximum obtainable deterministic revenue with capacity unconstrained:

[19]As explained by the authors (1976), peak load pricing may also be optimal in the absence of rationing costs if multiple plant types are available for use. This arises because such a diverse technology gives rise to a system-wide marginal cost function of supply which is upward sloping, with consequent higher marginal costs in peak periods.

[20]Harvey Averch and Leland Johnson presented the classical analysis of rate-of-return regulation. Elizabeth Bailey extended it to encompass peak loads.

[21]See Bailey for information on this model under certainty.

$$(19) \qquad R^* = \underset{p \geq 0}{\text{Max}} \sum_{i=1}^{2} (p_i - b) D_i(p_i)$$

Letting p^* be the optimal pricing solution to (19), it is straightforward to show[22] that if

$$(20) \qquad R^*/s \geq \underset{i \in \{1,2\}}{\text{Max}} [D_i(p_i^*) + u_{2i}]$$

then the regulated monopolist will set price equal to p^* and $Q^* = R^*/s$, thus achieving equality in the regulatory constraint. This being so, we see from (20) and the definition of u_{2i} that a reliability level of unity will be optimal for such a monopolist.

As an example, consider the demand curves of Section III with $b = 5$ and $\beta = 10$. It is easily verified that (20) is satisfied for any s, β, and γ satisfying $s/\beta \leq 7$ and $\gamma \leq 30$. For the case $s/\beta = 1.1$ and for any $\gamma \leq 30$, the solution to (15)–(17) is $p_1 = 142.5$, $p_2 = 58.5$, and $Q = 431.4$. For this case, therefore, Q^* is a whopping 5-fold multiple of even the welfare-maximizing optimal capacity. Thus, the statement in Meyer that "apparent excessive investment may be simply an optimal economic decision when the firm must meet high reliability standards" (1975, p. 336) masks the very important point that not only Averch-Johnson effects and reliability but also "gold plating" and reliability work in the same direction. If this is not curbed by some form of reliability constraint, very large excess capacity could be the result.[23] In particular, the above analysis suggests that regulation of rate of return should be coupled with some form of either specific reliability level constraints or maximal levels. Specifying minimal acceptable reliability levels alone may result in these being satisfied with a vengeance (i.e., more capacity is installed than that required to give a reliability level of unity). Indeed, the above example provides an illustration of gold plating if ever there was one.

Evidently, a more detailed analysis of reliability and regulation is required before proceeding to general conclusions. What does seem warranted at this stage is a certain pessimism regarding the efficacy of rate-of-return regulation if the reliability level is unregulated.[24] Regulation of reliability, however, raises extremely difficult questions regarding optimal reliability levels, depending as they do on estimating and apportioning the rationing and social costs of excess demand. Even more perplexing is the question of viable regulatory policies when the reliability epicycle is added to an already ponderous ptolemaic regulatory system. Nonetheless, answers to these questions are crucial if regulation is to mean something more than ill-conceived tampering.

[22]Let $Q^* = R^*/s$ and suppose (20) holds. If the optimal capacity Q is assumed less than Q^*, we have

$$E\{\pi(p, Q, u)\} \leq (s - \beta)Q^* = E\{\pi(p^*, Q^*, u)\}$$

where the first inequality holds since any optimal solution (p, Q) must satisfy (16), the strict inequality follows from $(s - \beta) > 0$, and the final equality holds by (20) and the definition of Q^*. Thus, for any feasible price p, Q cannot be optimal. Suppose, likewise, that $Q > Q^*$. Then

$$E\{\pi(p, Q, \tilde{u})\} = E\{\text{Rev}(p, Q, u)\} - \beta Q < R^* - \beta Q^* = E\{\pi(p^*, Q^*, \tilde{u})\},$$

where the inequality follows by definition of R^* as the maximum obtainable revenue. We see again that Q is nonoptimal. Clearly Q^* is the optimal capacity and, since (20) holds, p^* is feasible and also optimal in (15)–(17).

[23]The recent paper by Michael Telson lends some empirical support to this as well. His analysis suggests that current levels of reliability are far higher than his ball park cost-benefit analysis would imply as being reasonable.

[24]The "used and useful" criterion, as mentioned by Fred M. Westfield in dealing with a similar problem of excessive capitalization under certainty, does not answer the question of what level of reliability is useful.

REFERENCES

H. Averch and L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1053–69.

E. E. Bailey, "Peak-Load Pricing under Regulatory Constraint," *J. Polit. Econ.*, July/Aug. 1972, *80*, 662–79.

——— **and L. J. White,** "Reversals in Peak

and Off-Peak Prices," *Bell J. Econ.*, Spring 1974, *5*, 75-92.

V. S. Bawa and D. S. Sibley, "Dynamic Behavior of a Firm Subject to Stochastic Regulatory Review," disc. paper no. 38, Bell Laboratories, Sept. 1975.

Patrick Billingsley, *Convergence of Probability Measures*, New York 1968.

G. Brown, Jr. and M. B. Johnson, "Public Utility Pricing and Output Under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119-28.

M. A. Crew and P. R. Kleindorfer, "Peak Load Pricing with a Diverse Technology," *Bell J. Econ.*, Spring 1976, *7*, 207-31.

——— and ———, *Public Utility Economics*, London forthcoming.

R. E. Dansby, "Welfare Optimal Peak-Load Pricing and Capacity Decisions with Time Varying Demand," disc. paper no. 39, Bell Laboratories, Nov. 1975.

R. A. Meyer, "Monopoly Pricing and Capacity Choice Under Uncertainty," *Amer. Econ. Rev.*, June 1975, *65*, 326-37.

———, "Risk-Efficient Monopoly Pricing for the Multiproduct Firm," *Quart. J. Econ.*, Aug. 1976, *90*, 461-74.

W. Y. Oi, "A Disneyland Dilemma: Two Part Tariffs for a Mickey Mouse Monopoly," *Quart. J. Econ.*, Feb. 1971, *85*, 77-96.

R. Sherman and M. Visscher, "Second Best Pricing with Stochastic Demand," *Amer. Econ. Rev.*, Mar. 1978, *68*, 41-53.

P. O. Steiner, "Peak-Loads and Efficient Pricing," *Quart. J. Econ.*, Nov. 1957, *71*, 585-610.

M. Telson, "The Economics of Alternative Levels of Reliability for Electric Power Generation Systems," *Bell J. Econ.*, Autumn 1975, *6*, 679-94.

R. Turvey, "Public Utility Pricing and Output Under Risk: Comment," *Amer. Econ. Rev.*, June 1970, *60*, 485-86.

M. Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand: Comment," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224-29.

F. M. Westfield, "Conspiracy and Regulation," *Amer. Econ. Rev.*, June 1965, *55*, 424-43.

O. E. Williamson, "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints," *Amer. Econ. Rev.*, Sept. 1966, *56*, 810-27.

Second Best Pricing with Stochastic Demand

By ROGER SHERMAN AND MICHAEL VISSCHER*

Ways to depart from marginal cost pricing to increase revenue and yet minimize the resulting misallocation of resources are well-accepted members of a growing family of constrained welfare-maximizing prescriptions (see William Baumol and David Bradford for a review of the literature). An important application is found in public utility pricing, where optimal peak and off-peak pricing arrangements (for example, Marcel Boiteux 1949, Peter Steiner, or Oliver Williamson) have been modified as needed to admit second best characteristics (see Boiteux 1956 and Herbert Mohring). Second best two-part tariff schemes provide another example of pricing rules modified to satisfy budget constraints (see Yen-Kwang Ng and Mendel Weisser). But in all of these examples, demand is assumed known with certainty. Little attention has been given to second best solutions when demand has a random element,[1] little at least in comparison to the number of authors who have stressed the crucial role of demand uncertainty in determining either ideal or monopoly prices (Gardner Brown and M. Bruce Johnson, Hayne Leland, Robert Meyer, Edwin Mills, and Visscher, among others). Moreover, when demand is stochastic, deficits also are more likely at expected welfare-maximizing prices (Brown-Johnson, Meyer, and Visscher), so the need to raise revenue through a second best solution then is even greater. Randomness in demand is likely to play a role in determining optimal prices, too, and so existing second best pricing rules may be faulty when they yield prices tied only to marginal cost and demand elasticities while ignoring demand uncertainty.[2]

Our aim is to develop second best pricing rules for the important case of a monopoly firm facing stochastic peak and off-peak demands.[3] In our model the service capacity and prices in the different periods must be chosen before actual demands are known. When the actual demands do appear, excess demand (or supply) may exist because both the peak and off-peak demands for the service are stochastic. We assume service can be supplied at constant short-run marginal cost up to the capacity limit, so cost is not stochastic. We show the consequences of efficient and inefficient nonprice rationing of the available supply of service in such cases where prices cannot be adjusted in response to the actual demands.

We begin by obtaining for later comparison the prices and capacity that would be chosen by an unregulated profit-maximizing monopolist. In turning to welfare-maximizing solutions we first assume (as did Brown-Johnson and also Meyer) that persons who value service most will receive it even when demand exceeds capacity at the unchanging price; that is, markets always are cleared efficiently despite the absence of a genuine market-clearing price. In this case optimal prices turn out to be equal to short-run marginal costs and generally

*University of Virginia and Ohio State University, respectively. We are grateful to Dennis Epple, Edward Prescott, and a referee for helpful comments. Financial support from the International Institute of Management and the University of Virginia Center for Advanced Studies is also acknowledged.

[1]Maurice Marchand (1973) has imposed a break-even constraint on producers of telephone service operating under conditions of risk, where the effect of risk could be converted into a deterioration in service quality in the form of greater delays in telephone service. Such a treatment does not cover the possibility that service might actually be denied some consumers when demand exceeds capacity, but it does lead to rationing inefficiency. On service reliability, see Michael Crew and Paul Kleindorfer.

[2]The U.S. Postal Service relied on a simple second best pricing rule, which ignored randomness, in its 1974 postal rate proposals. See Postal Rate Commission Docket R74-1.

[3]Of course this model could easily be applied to a multiproduct firm without the peak load context, and it is also a simple matter to generalize to *n* products that share the same capacity (see Meyer).

do not produce sufficient revenue to cover costs. So by imposing a break-even constraint we can see how risk will influence optimal second-best prices. We find that both expected profit-maximizing and constrained expected welfare-maximizing solutions must now contain stochastic elements. The solutions are related to one another in a simple manner, however, analogous to the relation exhibited between profit-maximizing and constrained welfare-maximizing solutions when demand is certain. Further, the constrained welfare-maximizing solution implies a risk of failure that may be more defensible than one imposed arbitrarily as in Meyer's chance-constrained optimum.

We also treat the possibility that markets are not automatically cleared efficiently, because rationing inefficiency is a natural consequence when price cannot always clear markets. From among those willing to pay the quoted price when demand exceeds capacity, for instance, service may be awarded first to claimants with the least willingness to pay, or it may be distributed randomly among claimants. The presence of such inefficient rationing complicates the marginal conditions that must be satisfied if optimal capacity and prices are to be found, and makes profit-maximizing incentives inadequate to reach the welfare-maximizing solution. For with inefficient rationing, higher prices are appropriate to help the market clear more efficiently, and optimal capacity will tend to be larger also, than in the monopoly case or the second best welfare-maximizing case where, by magical assumption, markets automatically clear efficiently without price adjustments.

I. The Unregulated Monopolist's Solution

Consider a peak load pricing problem involving n services such as units of transportation consumed in n different time periods. The quantity per time period can never exceed the rate determined by capacity z, which has a cost of β for each unit of service it is capable of producing over the demand cycle. Each unit actually produced also requires an operating cost of b. In the ith period, demand is $X_i(P_i) + u_i$, where P_i is price, $X_i(P_i)$ is expected quantity demanded at price P_i, and u_i is a random variable with mean zero.[4] The function $F_i(\bar{u}_i)$ is a cumulative distribution and $f_i(u_i)$ is a density function such that $F_i(\bar{u}_i) = \int_{-\infty}^{\bar{u}_i} f_i(u_i)\,du_i$. The ith period lasts a fraction α_i of the total time interval for the demand cycle under consideration (for example, day, month, year), and n periods account for the total time interval so $\sum_{i=1}^{n} \alpha_i = 1$. We assume these demands and distributions for different periods are independent of each other,[5] and we consider a horizon of only one time interval.

Now suppose there is one monopolistic producer whose object is to choose capacity z^* as well as n prices P_i^*, where $i = 1, \ldots, n$ time periods over the demand cycle, to maximize expected profit. The starred variables represent the producer's chosen values, to distinguish them from other values. The expected profit of this monopoly firm can be represented as

$$(1)\quad E(\pi) = \sum_{i=1}^{n} \alpha_i (P_i^* - b)\{X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} [X_i(P_i^*) + u_i - z^*] f_i(u_i)du_i\} - \beta z^*$$

Expected profit without any capacity cost or capacity limit would simply be the sum of contributions to profits from all periods, $\sum_i \alpha_i(P_i^* - b)X_i(P_i^*)$. But because quantity is limited by costly capacity we must also take account of foregone profit when de-

[4]The welfare-maximizing rules developed in Section II (efficient nonprice rationing) and in Section IV (completely inefficient nonprice rationing) were derived for the case of multiplicative uncertainty and found to differ from those in the text only in unimportant ways such as the limits of integration and the form for expected excess demand. However, as Carlton has neatly shown, in the special case of *random* rationing of available supply whenever excess demand appears, the unconstrained welfare maximum is a break-even solution if the demand uncertainty is multiplicative. In that case there is no need for a second best, zero profit constrained analysis.

[5]Treatment of interdependence of demands in the different periods is foregone here; to obtain further results would require much greater complication of the model.

mand cannot be satisfied because $X_i(P_i^*) + u_i$ exceeds z^*. Moreover we must subtract the cost of capacity βz^*.

We must introduce a limitation in this model. The error term u_i can never take a value so negative that $-u_i$ exceeds $X_i(P_i^*)$. This assumption that $X_i(P_i^*) + u_i \geq 0$ ensures that quantity demanded is never negative.[6] As a practical matter such a requirement is certainly plausible, because at a given price the fluctuations in quantity we ordinarily can expect are small in proportion to average quantity. This limitation that randomness not overwhelm other aspects of the problem will be crucially important when we consider effects of inefficient nonprice rationing in Section III.

An expected profit-maximizing monopolist operates where the derivatives of (1) with respect to capacity and to the n prices are all equal to zero. Taking these derivatives and setting them equal to zero we obtain

$$(2)\quad \frac{\partial E(\pi)}{\partial P_i^*} = \alpha_i\{X_i(P_i^*) - \int_{z^*-X_i(P^*)}^{\infty} [X_i(P_i^*) + u_i - z^*] f_i(u_i)du_i + (P_i^* - b)[X_i'(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} X_i'(P_i^*) f_i(u_i)du_i]\} = 0 \qquad i = 1,\ldots,n$$

$$(3)\quad \frac{\partial E(\pi)}{\partial z^*} = \sum_{i=1}^{n} \alpha_i(P_i^* - b) \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)du_i - \beta = 0$$

Letting demand elasticities in all periods be represented as

$$(4)\quad \eta_i = \frac{-P_iX_i'(P_i)}{X_i(P_i)} \qquad i = 1,2,\ldots,n$$

[6]This assumption is quite essential. The alternative adopted by Meyer of constraining quantity to be no less than zero is an improvement over earlier analyses that ignored the potential problem of negative quantities, but it has the drawback that if such a constraint is effective, the distribution of u_i effectively will be truncated. To that extent $E(u_i)$ will depend on P_i^*. Having $E(u_i)$ depend on P_i^* would greatly complicate the analysis.

we obtain from (2) the implicit pricing rule:

$$(5)\quad \frac{(P_i^* - b)F_i(z - X_i(P_i^*))}{P_i^*} = \frac{1}{\eta_i} \cdot [1 - E(ed_i)/X_i(P_i^*)] \qquad i = 1,2,\ldots,n$$

where

$$(6)\quad E(ed_i) = \int_{z^*-X_i(P_i^*)}^{\infty} \cdot [X_i(P_i^*) + u_i - z^*] f_i(u_i)du_i$$

is expected excess demand. Setting $\partial E(\pi)/\partial z^* = 0$ in (3) yields the capacity condition

$$(7)\quad \sum_{i=1}^{n} \alpha_i(P_i^* - b) \cdot [1 - F_i(z^* - X_i(P_i^*))] = \beta$$

The term $1 - F(z^* - X_i(P_i^*))$ in the capacity condition represents the probability of excess demand; thus equation (7) clearly states that at expected profit-maximizing z^*, the marginal expected contribution to net revenue (from increasing capacity and allowing an additional unit to be sold at prices P_i^*, $i = 1,\ldots,n$ when excess demand occurs) is just equal to the marginal cost of extra capacity. The pricing rule in equation (5) is less transparent and it merits closer examination.

Equation (5) suggests that if the monopolist can choose inverse elasticity prices at the optimal z^* without inducing excess demand, those prices are the most profitable. But if excess demand is expected at those prices the profit-maximizing monopolist should forsake inverse elasticity pricing for a higher[7] price that would come closer to clearing the market. At the higher price, $F_i(\ )$ is larger and $E(ed_i)$ is smaller.

II. Constrained Expected Welfare Maximization Assuming that Markets Clear Efficiently

Now let us seek criteria for maximizing expected welfare while requiring the firm to break even on its operations. As in the monopoly case, we pose the problem in a

[7]Presuming $\partial^2X(P)/\partial P^2$ is not overwhelmingly positive, so the profit function remains concave.

partial equilibrium setting, and assume that Pareto conditions are always satisfied elsewhere in the economy. We also ignore income effects and assume risk neutrality so that expected consumer's surplus can serve as an indicator of welfare. The objective is to choose prices P_i^* for $i = 1, 2, \ldots, n$, and capacity z^*, to maximize a welfare function equal to the expected consumer's surplus plus revenue (i.e., willingness to pay) less expected cost (i.e., variable cost plus capacity cost), subject to the constraint that total revenue will just equal total cost. In treating the welfare-maximizing problem an assumption will be needed about the willingness to pay of those who are actually served whenever quantity demanded exceeds capacity. Such a question was of no interest to the expected profit-maximizing monopolist because profit, unlike consumer's surplus, is not affected by the assumption.

In this section we assume that the service produced is always distributed efficiently in the sense that consumers who value it most are the ones who receive the service, whether or not money price is high enough to ensure that result. This "efficient nonprice rationing" assumption was introduced by Brown and Johnson and also relied upon by Meyer. The assumption is interesting. When demand is stochastic and there is no break-even constraint, it leads to welfare-maximizing prices that equal short-run marginal costs (Brown-Johnson). But these prices do not yield enough revenue to cover total cost (although there are no economies of scale in production) and so a balanced budget-constrained second best welfare analysis is relevant. In Sections III and IV we explore the implications of relaxing this efficient nonprice rationing assumption.

There is an important contrast to be anticipated as we move from certainty to the stochastic demand case. At the optimum solution under certainty, excess demand would never occur. Indeed, with the cost function assumed here, the budget constraint always could be satisfied under certainty with efficient market-clearing prices (as, for example, in Williamson), and the budget constraint therefore would be inessential, or redundant. When demand is random, though, we have no assurance that any price or capacity level chosen before demand is known will clear the market once demand is revealed. And as we just noted, Brown and Johnson have shown that if efficient nonprice rationing is assumed in this situation, the welfare-maximizing set of prices and capacity will not allow the producer to break even. For if prices are set high enough to break even on average, capacity will go unutilized whenever demand is unusually low. Some consumers who valued the service above its marginal cost would be denied service then, and the outcome would be inefficient. To avoid such a result first best prices are set at short-run marginal cost, lower than break-even prices. But that means a break-even constraint will be binding in the stochastic demand problem even though it was not binding in the certainty model.

Let us now add a stochastic element to the demand function and derive a solution to the balanced budget constrained, second best problem. Assume that, as in the expected profit-maximizing model, demand in period i is given by the function $X_i(P_i) + u_i$ where u_i is distributed with density function $f_i(u_i)$ and mean zero. Again, period i lasts a fraction α_i of the demand cycle, and z^* and the P_i^*'s are unchanging values to be chosen and held at the same levels regardless what actual values of u_i occur. Total expected consumer's surplus and revenue is

$$\sum_{i=1}^{n} \alpha_i \left\{ \int_{\infty}^{\infty} f_i(u_i) \cdot \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i] dP_i du_i + P_i^* X_i(P_i^*) \right\}$$

less the expected loss in both consumer's surplus and revenue because some consumers are not served (in this case those who value service least) when quantity demanded exceeds capacity:

$$\sum_{i=1}^{n} \alpha_i \left\{ \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i) \cdot \right.$$

$$\int_{P_i^*}^{X_i^{-1}(z^*-u_i)} [X_i(P_i) + u_i - z^*]dP_i du_i +$$

$$\left. P_i^* \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,[X_i(P_i^*) + u_i - z^*]du_i \right\}$$

Variable costs are

$$\sum_{i=1}^{n} \alpha_i b \left\{X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \cdot [X_i(P_i^*) + u_i - z^*]du_i\right\}$$

and capacity cost is βz. The balanced budget constraint requires that

$$\sum_{i=1}^{n} \alpha_i(P_i^* - b)\left\{X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \cdot [X_i(P_i^*) + u_i - z^*]du_i\right\} = \beta z^*$$

We can thus construct a second best problem, maximizing expected consumer's surplus plus revenue less variable and capacity costs subject to the break-even constraint, by forming the Lagrangian

$$(8)\quad L(P_i^*, z^*, \lambda) = \sum_{i=1}^{n} \alpha_i \left\{\int_{-\infty}^{\infty} f_i(u_i) \cdot \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i]\,dP_i du_i + P_i^* X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) \int_{P_i^*}^{X_i^{-1}(z^*-u_i)} \cdot [X_i(P_i) + u_i - z^*]\,dP_i du_i - P_i^* \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i - b[X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i - \beta z^*\right\}$$

$$+ \lambda \left\{\sum_{i=1}^{n} \alpha_i(P_i^* - b)[X_i(P_i^*) - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*]\,du_i] - \beta z^*\right\}$$

Differentiating the Lagrangian with respect to each P_i^* and z^*, and setting results equal to zero, we have

$$(9)\quad \frac{\partial L}{\partial P_i^*} = \alpha_i\left\{\lambda X_i(P_i^*) - \lambda\int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i(P_i^*) + u_i - z^*]\,du_i + (1 + \lambda)(P_i^* - b)X_i'(P_i^*)F(z^* - X_i(P_i^*))\right\} = 0$$

$$(10)\quad \frac{\partial L}{\partial z^*} = \sum_{i=1}^{n} \alpha_i\left\{\int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i^{-1}(z^* - u_i) - P_i^*]\,du_i + (P^* - b)\int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,du_i\right\} - \beta + \lambda\left\{\sum_{i=1}^{n} \alpha_i(P_i^* - b) \cdot \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\,du_i - \beta\right\} = 0$$

Remember from (4) above our definition of demand elasticities η_i, and from (6) the definition of expected excess demand $E(ed_i)$. We can obtain from (9) optimum pricing rules in the form

$$(11)\quad \frac{(P_i^* - b)F(z^* - X_i(P_i^*))}{P_i^*} = \frac{\lambda}{1 + \lambda} \cdot \frac{1}{\eta_i}\left[1 - \frac{E(ed_i)}{X_i(P_i^*)}\right] \qquad i = 1, \ldots, n$$

Condition (11) is a constrained expected welfare-maximizing counterpart to the expected profit-maximizing implicit pricing rule in equation (5). Comparison of equation (11) with equation (5) shows that the only difference in form between the second best expected welfare-maximizing profit margin $(P_i^* - b)/P_i^*$ and the expected profit-maximizing profit margin is the constant term $\lambda/(1 + \lambda) < 1$ at the right-hand side of (11).

The relation between the pricing rules for monopoly (5) and for welfare (11) goals involves only a constant on the right-hand side as in the relation without risk illustrated by Baumol and Bradford. This is not surprising when one realizes that short-run marginal cost b is the first best price here if

nonprice rationing is assumed to be efficient. Prices are not used to ensure that service is alloted to those with the highest willingness to pay; some nonprice mechanism accomplishes that by assumption, while prices alone turn away only those unwilling to pay marginal operating costs. Thus prices serve the same function in this case for both the monopoly profit maximizer and unconstrained welfare maximizer, and that purpose is to raise revenue while turning away as few customers as possible. So in the presence of risk the relation between prices for the unconstrained and the constrained monopolist can be analogous to the comparable relation ignoring risk; the constrained welfare maximizer raises price above marginal cost by the greatest amount in the same periods as would an expected profit maximizer, but only by an amount sufficient to break even. When properly constrained the monopolist's incentives therefore can still be efficacious for maximizing welfare, once stochastic terms are properly taken into account.

From (10) we can obtain the requirement for optimum capacity.

$$(12)\quad \sum_{i=1}^{n} \alpha_i \left\{ \frac{1}{1+\lambda} \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i^{-1}(z^* - u_i) - b]\, du_i + \left(\frac{\lambda}{1+\lambda}\right) \cdot (P_i^* - b)(1 - F[z^* - X_i(P_i^*)]) \right\} = \beta$$

Optimum capacity choice under stochastic demand is analogous to that under nonstochastic demand. When demand is certain it is known whether excess demand will exist in period i at given P_i^*, z^* choices, while if demand is stochastic, excess demand will exist only with some probability. Additional capacity yields benefits in either case only if excess demand exists. The marginal benefit from extra capacity would be known exactly if demand were nonstochastic, but since demand is stochastic here, the left-hand side of (12) represents *expected* marginal benefit. The benefit appears as added consumer's surplus and added net revenues which contribute to welfare directly and further help to satisfy the break-even constraint. If demand were certain the break-even constraint would not be binding ($\lambda = 0$) as long as the P_i^* and z^* were optimally chosen because, given our cost function, total revenue would equal total cost. However, the break-even constraint is binding ($\lambda > 0$) when P_i^*, z^* are chosen optimally in the stochastic demand model, for without the constraint the firm would lose money as Brown and Johnson have shown.

Under stochastic demand the welfare-maximizing capacity rule in (12) can be related to the profit-maximizing rule in (7). Notice that as the amount of net operating revenue needed to break even increases, the shadow price on the budget constraint λ increases. Marginal net revenue then becomes the important concern for the welfare maximizer when deciding whether to alter capacity size, just as it is for the profit maximizer; as λ becomes larger, equation (12) (the second best capacity solution) becomes more like equation (7) (the profit-maximizing capacity solution). And (12) also yields implicitly a reliability of service that is optimal given the break-even constraint, and thus does not have to be arbitrarily imposed (see Meyer).

III. The Consequences of Inefficient Nonprice Rationing

We now examine effects on price and capacity of an alternative and more reasonable rationing assumption, namely that a genuine market-clearing price is needed for efficient rationing. When nonprice rationing is inefficient, price has a role other than merely raising revenue; price can deny service to claimants with low willingness to pay so those who value the service more can be supplied from existing capacity. We introduce inefficiency of nonprice rationing into the model of this section simply by assuming that when demand exceeds capacity

at the price chosen, the least efficient possible rationing results.[8]

The social loss from excess demand in cases of inefficiently cleared markets will now be higher than in the previous section, where efficient market clearing was always assumed. We should expect welfare-maximizing pricing rules to reflect the greater loss by requiring either higher prices or larger capacities, or both, so excess demand will be less likely. We show such higher optimal prices and capacity are assured under our assumptions, and explain the unusual case in which rationing inefficiency might require lower prices and smaller capacity.

We shall set out our general line of argument and then evaluate it using, with appropriate modifications, the more specific functional forms already introduced. Let $\phi = \phi(P_1, P_2, \ldots, P_n, z)$ be expected consumer's surplus and let $\pi = \pi(P_1, P_2, \ldots, P_n, z)$ be expected producer's surplus. To maximize $\phi + \pi$ by choice of $P_1, P_2, \ldots, P_n, z$, subject to $\pi = 0$, we construct the Lagrangian

$$L(P_1, P_2, \ldots, P_n, z, \lambda) = \phi + (1 + \lambda)\pi$$

Setting derivatives equal to zero (and ignoring the possibility of corner solutions), we have

$$\frac{\partial L}{\partial P_i} = \frac{\partial \phi}{\partial P_i} + (1 + \lambda)\frac{\partial \pi}{\partial P_i} = 0$$

$$i = 1, 2, \ldots, n$$

and $$\frac{\partial L}{\partial z} = \frac{\partial \phi}{\partial z} + (1 + \lambda)\frac{\partial \pi}{\partial z} = 0$$

indicating that at the constrained optimum

$$\left.\frac{dz}{dP_i}\right|_{d\phi = 0} = \left.\frac{dz}{dP_i}\right|_{d\pi = 0} \qquad i = 1, 2, \ldots, n$$

and $$\left.\frac{dP_j}{dP_i}\right|_{d\phi = 0} = \left.\frac{dP_j}{dP_i}\right|_{d\pi = 0} \qquad i, j = 1, 2, \ldots, n$$

[8]Of course any fraction of this potential consumer's surplus might be realized, depending on the efficiency of nonprice rationing. For a description of extreme and intermediate representations of inefficient nonprice rationing, see Gene Mumy and Steve Hanke.

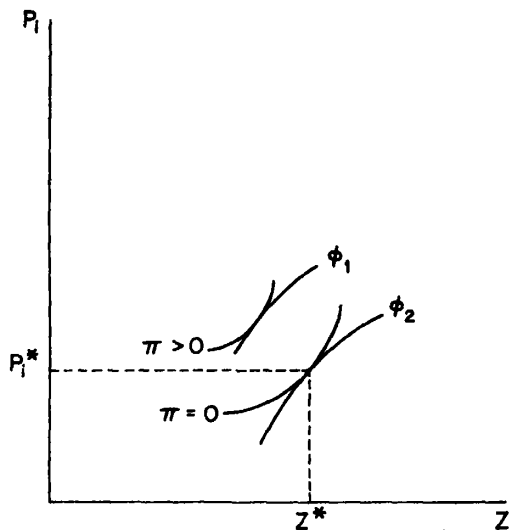



FIGURE 1

Thus, the same changes in price and capacity that will keep expected consumer's surplus constant will also keep expected producer's surplus constant.

The situation is illustrated in Figure 1 for a single P_i and z. The profit function is assumed convex. Iso-profit contours labeled π are upward sloping to the right, because higher prices should raise revenues in the relevant region and make expenditure on capacity necessary to satisfy the budget constraint. Iso-consumer's surplus contours labeled ϕ also are upward sloping to the right because higher prices can be introduced while expected consumer's surplus remains constant only if capacity is also increased.

Recall now that the efficiency of nonprice rationing can affect expected consumer's surplus, but not profit or the monopoly solution (equations (5) and (7)), and so it will not affect the profit contours in Figure 1. Since the iso-profit contours will not change and we consider throughout the same zero profit contour, then to prove that inefficient rationing will lead to higher prices and larger capacity it is sufficient to show that the slope of the iso-consumer's surplus contours will decrease with less effi-

cient rationing. For if the iso-ϕ contour is to have a lower slope its tangency with the zero iso-π contour must occur at greater P_i and z. We can show that this condition is likely to be satisfied, although the possibility also will exist that optimum prices and capacity are *smaller* when rationing is inefficient.

The effect of rationing inefficiency can be demonstrated by comparing the derivatives of consumer's surplus with respect to price and capacity when nonprice rationing is efficient, as in Section II, with those same derivatives when rationing is inefficient. Rationing is most inefficient if, whenever demand exceeds the available capacity, sales are made first to those potential claimants with the least willingness to pay.[9] We now return to our assumed functional forms and make these comparisons.

With efficient rationing, expected consumer's surplus as given in (8) will be

$$\phi_{ER} = \sum_{i=1}^{n} \alpha_i \left\{ \int_{-\infty}^{\infty} f_i(u_i) \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i]\, dP_i\, du_i - \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i) \int_{*P_i}^{X^{-1}(z^* - u_i)} [X_i(P_i) + u_i - z^*]\, dP_i du_i \right\}$$

where the second term represents lost expected consumer's surplus due to the chance of drawing a value of u_i so large that excess demand results. Notice we integrated over values of P_i from P_i^* to $X_i^{-1}(z^* - u_i)$, because claimants with the lowest demand prices were assumed to be denied service whenever excess demand appeared under the efficient nonprice rationing rule of Section II. Taking derivatives we have

$$(13) \quad \frac{\partial \phi_{ER}}{\partial P_i^*} = \alpha_i \left\{ - \int_{-\infty}^{\infty} f_i(u_i) [X_i(P_i^*) + u_i] du_i + \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i)[X_i(P_i^*) + u_i - z^*] du_i \right\} < 0$$

if $z^* > 0$, and

$$(14) \quad \frac{\partial \phi_{ER}}{\partial z^*} = \sum_{i=1}^{n} \alpha_i \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i) [X_i^{-1}(z^* - u_i) - P_i^*] du_i > 0$$

The slope of the iso-consumer's surplus contour is the ratio of $-\partial\phi_{ER}/\partial P_i^*$ to $\partial\phi_{ER}/\partial z^*$ and it is clearly positive as drawn in Figure 1.

Expected consumer's surplus when nonprice rationing is so inefficient that those with the least willingness to pay are served first is given by

$$\phi_{IR} = \sum_{i=1}^{n} \alpha_i \left\{ \int_{-\infty}^{\infty} f_i(u_i) \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i]\, dP_i\, du_i - \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i) \int_{X_i^{-1}[X_i(P_i^*) - z^*]}^{X^{-1}(-u)} [X_i(P_i) + u_i]\, dP_i\, du_i - \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i)[X_i^{-1}(X_i(P_i^*) - z^*) - P_i^*] \cdot [X_i(P_i^*) + u_i - z^*] du_i \right\}$$

where the last two terms now represent the lost expected consumer's surplus due to the possibility of excess demand. Notice this loss reflects the fact that claimants with demand prices from $X_i^{-1}[X(P_i^*) - z^*]$ to $x_i^{-1}(-u_i)$, the *highest* demand prices, are the ones now denied service when there is excess demand, because they are the ones to be counted in the welfare loss when rationing is most inefficient. The derivatives are

[9] This extremely inefficient nonprice rationing is not unreasonable to assume if queues occur. To the extent willingness to pay is related to the opportunity cost of one's time, persons with the lowest willingness to pay might also be the most willing to wait in line. See D. Nichols, E. Smolensky, and T. N. Tideman.

$$(15)\quad \frac{\partial \phi_{IR}}{\partial P_i^*} = \alpha_i \Big\{ -\int_{-\infty}^{\infty} f_i(u_i) [X_i(P_i^*) + u_i]\, du_i + \int_{z^*-X(P_i^*)}^{\infty} f_i(u_i) [X_i(P_i^*) + u_i - z^*]\, du_i - \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) X_i'(P_i^*)[X_i^{-1}(X_i(P_i^*) - z^*) - P_i^*]\, du_i \Big\} < 0$$

$$(16)\quad \frac{\partial \phi_{IR}}{\partial z^*} = \sum_{i=1}^{n} \alpha_i \Big\{ \int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i) [X_i^{-1}(X_i(P_i^*) - z^*) - P_i^*]\, du_i \Big\} > 0$$

The slope of the iso-consumer's surplus contour is again positive. But the slope of this contour is less when nonprice rationing is less efficient at P_i^*, z^* values such that tangency is maintained with the zero profit contour.

If we consider (13) and (15), it is easy to see that (15) contains an added positive term[10] which, with less efficient rationing at the same P_i^* and z^* values, would tend to make $-\partial\phi/\partial P_i$ smaller. This means that in the numerator of the slope of the iso-consumer's surplus contour,

$$(16')\quad \left.\frac{dz}{dP_i^*}\right|_{d\phi=0} = \frac{-\partial\phi}{\partial P_i} \Big/ \frac{\partial \phi}{\partial z}$$

we can expect $-\partial\phi_{ER}/\partial P_i > -\partial\phi_{IR}/\partial P_i$. By comparing (14) and (16) we can see it is likely that the denominator will be larger with less efficient rationing, or $\partial\phi_{ER}/\partial z < \partial\phi_{IR}/\partial z$. Together with the effect in the numerator this result would confirm that the slope given by (16′) is smaller as nonprice rationing is less efficient, and thus that larger P_i and z are then warranted. But the result for the denominator from (14) versus (16) is not assured unambiguously. We illustrate the relevant magnitudes in Figure 2.

[10]The third term in (15) is subtracted, but because it contains $X_i'(P_i^*)$ the term is negative and the net effect therefore is positive.

Equation (16) gives the expected incremental consumer's surplus from added capacity when rationing is inefficient. The extra consumer's surplus in period i is the sum of the difference between the highest valuation now served when excess demand appears, the valuation represented by $X_i^{-1}(X_i(P_i^*) - z^*)$ in Figure 2, and price P_i^*. The value of extra capacity stems from the *highest* valuation served during excess demand because capacity is filled by consumers with the *least* willingness to pay when rationing is inefficient, and added capacity thus allows the person with the next highest valuation to be served. Equation (14), on the other hand, displays the value of added capacity to expected consumer's surplus when rationing is efficient. Marginal consumer's surplus in period i then is the expected difference between the lowest valuation now served when excess demand appears, $X_i^{-1}(z^* - u_i)$, the price P_i^*. The lowest valuation $X_i^{-1}(z^* - u_i)$ varies with u_i, and values of u_i which are high and very high ($\hat{u}_i$ and $\hat{\hat{u}}_i$) are shown in Figure 2. The *lowest* valuation served determines the marginal values of capacity in this efficient rationing case, because existing capacity is already going to customers with the highest willingness to pay.

Now it can be seen in Figure 2 that the value of added capacity when rationing is efficient is greater than its value when rationing is inefficient only if some extremely high draws of u_i are likely. That is, $\partial\phi_{ER}/\partial z^* > \partial\phi_{IR}/\partial z^*$ only if draws of u_i will satisfy

$$X_i^{-1}(X_i(P_i^*) - z^*) < X_i^{-1}(z^* - u_i)$$

or

$$X_i(P_i^*) + u_i > 2z^*$$

in enough periods to make (14) exceed (16). We have already chosen to rule out the possibility of large negative u_i values such that $X_i(P_i^*) + u_i < 0$. If we similarly rule out large positive values of u_i that satisfy the roughly symmetrical requirement, $X_i(P_i^*)$

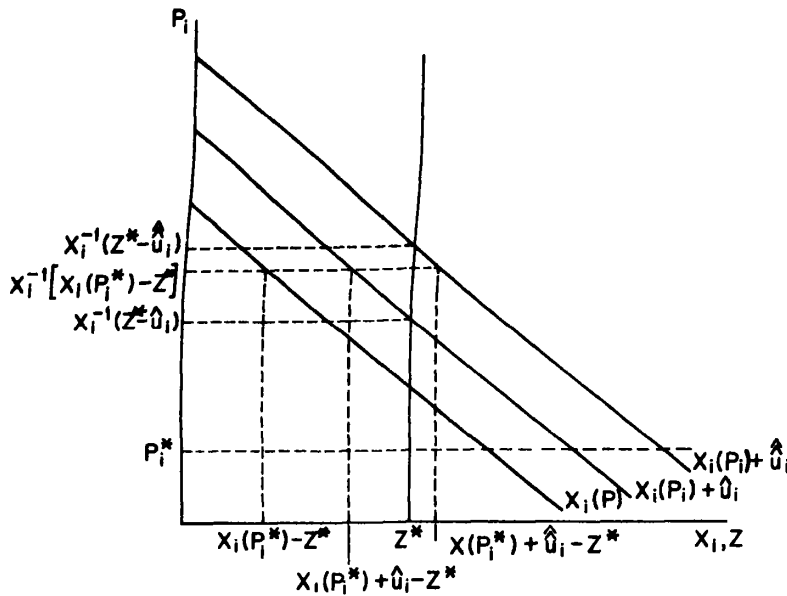



FIGURE 2

$+ u_i > 2z^*$, we shall always have (16) exceed (14). As a practical matter it does not seem that we restrict the model much by ruling out the possibility that random elements can overwhelm other aspects of the problem. And then rationing inefficiency will always require larger optimal prices and capacity.

IV. Constrained Expected Welfare Maximization When Markets are Not Cleared Efficiently

It remains to derive implicit choice rules from optimum prices and capacity when nonprice rationing is inefficient. Let us form a new Lagrangian from (8) by substituting the expression for consumer's surplus under the least efficient rationing assumption as set out in Section III.[11] By differentiating this modified Lagrangian with respect to z^* and P_i^*, and setting the results equal to zero we obtain:

[11]We represent the capacity constraint here by separate treatment of its effects when it is binding, aided by nonprice rationing assumptions, and we then rely on an explicit constraint only to represent the break-even requirement.

$$(17)\quad \left(\frac{P_i^* - b}{P_i^*}\right) F_i[z^* - X_i(P_i^*)] = \frac{\lambda}{1+\lambda} \cdot \frac{1}{\eta_i}\left[1 - \frac{E(ed_i)}{X_i(P_i^*)}\right] + \left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{X_i^{-1}[X_i(P_i^*) - z^*] - P_i^*}{P_i^*}\right) \cdot (1 - F[z^* - X_i(P_i^*)])$$

$$(18)\quad \sum_{i=1}^{n} \alpha_i \left\{ \left(\frac{1}{1+\lambda}\right) \int_{z^* - X_i(P_i^*)}^{\infty} f_i(u_i) (X_i^{-1}[X_i(P_i^*) - z^*] - b) du_i + \left(\frac{\lambda}{1+\lambda}\right)(P_i^* - b) \cdot (1 - F_i[z^* - X_i(P_i^*)]) \right\} = \beta$$

When rationing was efficient, the second best welfare-maximizing pricing rule differed from the expected profit-maximizing rule only by a constant term (compare (11) with (5)). But that is no longer the only difference if rationing in the presence of excess demand first serves claimants with the *least* willingness to pay. There is now a second

term in the right-hand side of (17) that reflects the impact of rationing inefficiency on the pricing rule. The manner in which this expression influences the optimum price is straightforward. The term $X_i^{-1}[X_i(P_i^*) - z^*] - P_i^*$ is the gain in consumer's surplus if price is raised slightly. When excess demand occurs and capacity is rationed to claimants with the least willingness to pay, this term represents the consumer's surplus lost by the last person turned away. Thus the second term on the right-hand side of (17) is equal to some positive multiple $1/(1 + \lambda)$ of the ratio of the net marginal gain in consumer's surplus from a higher price (if there is excess demand) to the value of the price chosen, all multiplied by the probability of excess demand occurring at P_i^*, z^* (the probability $1 - F[z^* - X_i(P_i^*)]$).

As expected, the ideal price is now higher than would be optimal for the same z^* under the efficient nonprice rationing assumption in (11) from Section II, since the new second term itself will raise the right-hand side of (17). Here price must serve a rationing function not required of it in Section II. After all, when price is the only means of assuring that service goes to highest willingness to pay claimants it should be nearer to the market-clearing level than when rationing always is efficient merely by assumption. Turning to the rules for optimal capacity, note that equation (18) is the inefficient rationing counterpart to equation (12) for efficient nonprice rationing. The only difference between them is in the marginal benefit from a unit of capacity, which depends on the $X_i^{-1}(z^* - u_i)$ term in (12) and the $X_i^{-1}(X_i(P_i^*) - z^*)$ term in (18). We ordinarily expect the expression in (18) will tend to be the larger, inviting larger capacity, because we expect $X_i(P_i^*) - z^* < z^* - u_i$ from the discussion in Section III (and Figure 2).

Let us briefly examine the effects of random rationing when excess demand exists. Random rationing gives each person who is willing to pay the going price an equal chance to be served by the available capacity, regardless of the magnitude of his marginal evaluation. Whenever there is random rationing of the available capacity with excess demand, the loss of consumer's surplus in the constrained welfare function of equation (8) will become

$$\int_{z^*-X_i(P_i^*)}^{\infty} f_i(u_i)\left(1 - \frac{z^*}{X_i^*(P_i^*) + u_i}\right) \cdot \int_{P_i^*}^{X_i^{-1}(-u)} [X_i(P_i) + u_i]dP_i du_i$$

Subtracting this loss from the Lagrangian in (8) rather than the efficient rationing loss will yield, on differentiating with respect to P_i and setting the result equal to zero,

$$(19)\quad \frac{P_i^* - b}{P_i^*} F_i[z^* - X_i(P_i^*)] = \left(\frac{\lambda}{1+\lambda}\right) \cdot \frac{1}{\eta_i}\left[1 - \frac{E(ed_i)}{X_i(P_i^*)}\right] + \left(\frac{1}{1+\lambda}\right)\frac{z^*}{P_i^*} \cdot \int_{z^*-X_i(P_i^*)}^{\infty} \frac{f_i(u_i)}{[X_i(P_i^*) + u_i]^2} \cdot \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i]dP_i du_i$$

This pricing rule is analogous to the one in equation (17), for the second term on the right-hand side of (19) is again positive; at the same z^* it will cause price to be higher than if efficient rationing was assumed as in pricing equation (11).

Now differentiate with respect to z^* the balanced budget constrained welfare function which was modified to allow random rationing. We can obtain similarly the capacity rule

$$(20)\quad \sum_{i=1}^{n} a_i\left\{\left(\frac{\lambda}{1+\lambda}\right)(P_i - b)(1 - F_i[z^* - X_i(P_i^*)]) + \left(\frac{1}{1+\lambda}\right)\int_{z^*-X_i^*(P_i)}^{\infty} \frac{f_i(u_i)}{X_i(P_i^*) + u_i} \cdot \int_{P_i^*}^{X_i^{-1}(-u_i)} [X_i(P_i) + u_i]dP_i du_i\right\} = \beta$$

Again, although the term is not exactly the same as in (18), the second term in brackets at the left-hand side in (20) will represent lost consumer's surplus due to excess demand, and in its presence, capacity will again tend to be increased beyond that given by the rule that assumes efficient nonprice rationing, equation (12).

V. Conclusions

Second best pricing rules have been developed here for a monopoly firm facing stochastic demands. We relied on the welfare function set out initially by Oliver Williamson, elaborated since to deal with risk by Brown and Johnson, and elaborated further to deal with risk plus peak and off-peak pricing by Meyer. If demand is uncertain and price must be set before demand in a particular period is known, some consumers will occasionally have to be denied service, because the total quantity actually demanded at the preannounced price can exceed the available capacity. When that happens the market may not clear, and without further information we have no basis for determining which consumers will be served. The expected profit-maximizing monopolist does not care how service is rationed among the consumers; the monopolist will choose prices taking into account marginal costs, demand elasticities, and also the probability distributions of demands. But when we seek to maximize expected welfare, the question about who is to be served first in cases of excess demand is crucial.

We first examined the case assumed by Brown and Johnson and also by Meyer in which consumers who value the service most are the ones who receive it when there is excess demand, even though there is no genuine market-clearing price to ensure that result. In this case we found rules for second best optimal prices which are just like the monopolist's rules, except for a constant term. So the contrast between monopoly and second best rules is comparable in this stochastic case to the well-known contrast obtained under certainty.

When we introduced inefficient nonprice rationing we found higher prices and a larger capacity are appropriate. We examined a case in which consumers who value the service least are the ones served first whenever there is excess demand, so the service is not distributed efficiently among claimants. The welfare-maximizing pricing rules can no longer be obtained merely by multiplying one side of the monopolist's pricing rules by a constant. Thus when demand uncertainty prevents price from ensuring efficient distribution of the service, a simple modification of the monopolist's solution no longer will yield the welfare-maximizing solution. The same was shown to be true when service is rationed randomly among consumers who are willing to pay the going price. This general difficulty arises because the profit-maximizing firm ignores consumer's surplus, and yet consumer's surplus will not reach an optimum level without special attention in the general case when nonprice rationing is inefficient.

The interesting question of pricing differently for different individuals, based in part on the regularity (or certainty) of their demands, has been treated by Boiteux (1951). He showed how the pooling of uncertain demands that are not perfectly and positively correlated can lead to welfare gains. This promising line of investigation was elaborated recently by Marchand (1974). We did not attempt ideal prices for individuals based on the consistency of their purchases. As noted by Meyer, however, an analysis of multiple demands can be interpreted as an analysis of customer groups, and so it would be possible to deal partly with the consistency of customers' usage in the model we have used.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer. Econ. Rev.*, Dec. 1962, *52*, 1052–69.

W. J. Baumol and D. F. Bradford, "Optimal Departures from Marginal Cost Prices," *Amer. Econ. Rev.*, June 1970, *60*, 265–83.

M. Boiteux, "La tarification au cout marginal et les demandes aléatoires," *Cahiers du Seminaire d'Econometrie*, 1951, *1*, 56-69.

———, "La tarification des demandes en pointe," *Revue Generale de l'Electricite*, 1949, 321-40; translated as "Peak-Load Pricing," *J. Bus. Univ. Chicago.*, Apr. 1960, *33*, 157-79.

———, "Sur la gestion des monopoles publics astreints à l'équilibre budgétaire," *Econometrica*, Jan. 1956, *24*, 22-40; translated as "On the Management of Public Monopolies Subject to Budgetary Constraints," *J. Econ. Theory*, Sept. 1971, *3*, 219-40.

G. Brown, Jr. and M. B. Johnson, "Public Utility Output and Pricing under Risk," *Amer. Econ. Rev.*, Mar. 1969, *59*, 119-28.

D. W. Carlton, "Market Behavior With Demand Uncertainty and Price Inflexibility," work. paper no. 179, Mass. Inst. Technology, June 1976.

M. A. Crew and P. R. Kleindorfer, "Reliability and Public Utility Pricing," *Amer. Econ. Rev.*, Mar. 1978, *68*, 31-40.

J. H. Drèze, "Some Postwar Contributions of French Economists to Theory and Public Policy," *Amer. Econ. Rev.*, June 1964, Suppl., *54*, 1-64.

M. G. Marchand, "The Economic Principles of Telephone Rates under a Budgetary Constraint," *Rev. Econ. Stud.*, Oct. 1973, *40*, 507-15.

———, "Pricing Power Supplied on an Interruptible Basis," *Euro. Econ. Rev.*, July 1974, *5*, 263-74.

R. A. Meyer, "Monopoly Pricing and Capacity Choice under Uncertainty," *Amer. Econ. Rev.*, June 1975, *65*, 326-37.

E. S. Mills, "Uncertainty and Price Theory," *Quart. J. Econ.*, Feb. 1959, *73*, 116-30.

H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *Amer. Econ. Rev.*, Sept. 1970, *60*, 693-705.

G. E. Mumy and S. H. Hanke, "Public Investment Criteria for Underpriced Public Products," *Amer. Econ. Rev.*, Sept. 1975, *65*, 712-20.

Y. Ng and M. Weisser, "Optimal Pricing with a Budget Constraint—The Case of the Two-Part Tariff," *Rev. Econ. Stud.*, July 1974, *41*, 337-45.

D. Nichols, E. Smolensky, and T. N. Tideman, "Discrimination by Waiting Time in Merit Goods," *Amer. Econ. Rev.*, June 1971, *61*, 312-23.

Roger Sherman, *The Economics of Industry*, Boston 1974.

M. Visscher, "Welfare-Maximizing Price and Output with Stochastic Demand," *Amer. Econ. Rev.*, Mar. 1973, *63*, 224-29.

O. E. Williamson, "Peak-Load Pricing and Optimal Capacity," *Amer. Econ. Rev.*, Sept. 1966, *56*, 810-27.

U.S. Postal Rate Commission, Docket R74-1, Tr. 2, pp. 201-445.

On the Optimality of Forward Markets

By ROBERT M. TOWNSEND*

Kenneth Arrow's seminal article on the role of securities in the optimal allocation of risk bearing provided a convenient framework in which problems involving choice under uncertainty could be analyzed. By extending the commodity space to include random states of nature, classic results on the existence and optimality of a competitive equilibrium were made applicable to uncertain situations. Yet many authors have commented on the existence of the small number of markets in which claims contingent on the realization of a state are actively traded. In particular the existence of futures or forward markets in which unconditional rather than contingent claims are traded is regarded by some as a phenomenon in need of an explanation, and by others as prima facie evidence of some inefficiency.

The purpose of this paper is to show that in some cases any equilibrium allocation resulting from the operation of competitive prestate noncontingent forward markets and competitive poststate spot markets is Pareto optimal, and that any Pareto optimal allocation can be supported as a competitive equilibrium of these markets with appropriate redistribution of endowments. These propositions turn on the fact that if equilibrium spot prices satisfy certain conditions then a restriction to the trading of forward contracts will not be constraining in an equilibrium; that is, agents can achieve precisely the same allocation with forward and spot markets as they could with markets in which claims could be traded for any commodity contingent on any state.

In his article Arrow stressed that in actual markets risk bearing is not allocated by the sale of claims against specific commodities but rather by the sale of securities payable in money, and he argued that any optimal allocation could be achieved with an elementary set of such securities. These Arrow-Debreu securities, as they have become known, suffice because their returns span the space of all possible returns. That is, any security whatever can be regarded as a bundle of these elementary securities, and, as has been noted by many authors, if an arbitrary set of securities spans the space of all possible returns, then such a set of securities is essentially equivalent to the set of Arrow-Debreu securities. In particular Steinar Ekern and Robert Wilson, and Roy Radner have argued that equities or shares may have the spanning property, and Steven Ross has made a similar argument for options. This paper shows that a forward contract may be viewed as a security whose return is the amount of the numeraire good (i.e., the price) for which it can be exchanged in the spot market of each state. Thus if the rank of the matrix of spot prices is equal to the number of states,[1] forward contracts also have the spanning property, and the results of this paper may be viewed as an extension of Arrow's results.[2]

*Carnegie-Mellon University. An earlier version of this paper appeared in my doctoral dissertation presented at the University of Minnesota, July 1975. I am especially indebted to Neil Wallace for his advice and encouragement. I also acknowledge the assistance of Paul Anderson and helpful comments from John Chipman, John Danforth, Hayne Leland, Stephen Salant, Leonard Shapiro, and the referee, but claim full responsibility for any errors or ambiguities. Financial support from the Board of Governors of the Federal Reserve System and from the Federal Reserve Bank of Minneapolis is gratefully acknowledged. The views expressed herein are solely my own and do not necessarily represent the views of the Bank or of the Federal Reserve System.

[1]This is not possible if there are more states than commodities.

[2]The extension however is not quite as immediate as it may first appear to be. Both Arrow's model and his result have been given a variety of interpretations. Arrow modeled a distribution economy in which money as an actual commodity plays a role. In contrast in the exchange economy of this paper there is no money per se. Thus Arrow's proof may not be ap-

This paper proceeds as follows. Section I presents the assumptions and technology of a pure exchange economy and describes the operation of two exchange regimes—complete prestate markets for contingent claims with no poststate spot markets and noncontingent forward markets with poststate spot markets. Section II formalizes the two welfare propositions given above and outlines their proofs. These results are then interpreted by way of some examples which clarify the nature of the spanning property. Section III presents an example which emphasizes the general equilibrium hedging property of forward contracts and provides further insight into the workings and welfare implication of forward markets when market structure is incomplete. Section IV presents some concluding remarks. Formal proofs are shown in the Appendix.

I. Description of the Model and Exchange Regimes

The model is a pure exchange economy with random endowments. There are I consumers, S mutually exclusive states of the world, and C commodities. Let π_s denote the probability that state s will occur with $0 < \pi_s < 1$. In this context endowments and consumption should be indexed by the consumer i ($i = 1, 2, \ldots, I$), state s ($s = 1, 2, \ldots, S$), and commodity c ($c = 1, 2, \ldots, C$) to which they pertain. Hence let Z_{isc} and C_{isc} denote the endowment and consumption, respectively, of consumer i in state s of commodity c, and let Z_{is} and C_{is} denote the associated C dimensional vectors. Each consumer i maximizes expected utility:

$$\sum_{s=1}^{S} \pi_s U^i(C_{is})$$

Each is assumed to be risk averse in that function $U^i(\ )$ is strictly concave.[3]

There are various possibilities for trade in the model. In what follows two exchange structures will be imposed exogenously and then compared. In the first exchange regime there are complete prestate markets for contingent claims. Trading in the markets for such claims takes place before random endowments are known. Also in the first regime there is no trading in spot markets subsequent to the realization of the state. In the prestate markets each consumer can issue or purchase contingent claims, where each claim entitles the holder to one unit of a specified commodity if a particular state occurs, and zero otherwise. Let X_{isc} denote the number of such unit claims on commodity c in state s held by consumer i after trading in the market for claims (with associated C dimensional vector X_{is}). That is, $(X_{isc} - Z_{isc})$ is the demand for such claims by consumer i in the market for claims. Let r_{sc} denote the price of a unit claim on commodity c in state s in terms of some abstract unit of account. Then the budget constraint for consumer i in the markets for claims is of the form

$$(1) \qquad \sum_{s=1}^{S} \sum_{c=1}^{C} r_{sc}(X_{isc} - Z_{isc}) = 0$$

After endowments are realized and some state is known to pertain, claims are honored so that $C_{is} = X_{is}$. In summary, in the first exchange regime consumer i maximizes

$$(2) \qquad \sum_{s=1}^{S} \pi_s U^i(X_{is})$$

with respect to $\{X_{isc}\}$ subject to (1) with each $X_{is} \geq 0$.[4] An equilibrium in the first exchange regime is a set of claim prices $\{r^*_{sc}\}$ and an allocation $\{X^*_{isc}\} i = 1, 2, \ldots, I$ such that $\{X^*_{isc}\}$ is maximizing for each consumer i and there is equality of the number of claims bought and sold for each state s and each commodity c, that is,

plied directly. Though the principal results of this paper are known by some, there does appear to be a need for a clarifying exposition. (The analogue of Arrow's theorem for an exchange economy is presented in Appendix B.)

[3]It is also assumed that $U^i(\ )$ is continuously differentiable with $U^i_C(0) = \infty$.

[4]$\{X_{isc}\}$ denotes the SC dimensional vector with elements X_{is}, $s = 1, 2, \ldots, S$; $c = 1, 2, \ldots, C$. This shorthand notation is used below for this and other variables if no ambiguity results.

$$(3) \quad \sum_{i=1}^{I} (X^{*}_{isc} - Z_{isc}) = 0$$

$$s = 1, 2, \ldots, S; \; c = 1, 2, \ldots, C$$

In the first exchange regime there was a restriction that there be no trading in spot markets subsequent to the realization of a state. But that restriction cannot be constraining in a competitive equilibrium. For let P_{sc} denote the spot price of commodity c in state s (with C dimensional vector P_s) where the Cth commodity is chosen as the numeraire. Now suppose that in state s some auctioneer calls out the vector P^{*}_{s} where

$$(4) \qquad P^{*}_{sc} = r^{*}_{sc}/r^{*}_{sC}$$

If trade is permitted, each consumer i is then confronted in the spot market of state s with the following problem: maximize $U^i(C_{is})$ with respect to C_{is} subject to the budget constraint

$$\sum_{c=1}^{C} P^{*}_{sc}(C_{isc} - X^{*}_{isc}) = 0 \qquad \text{with } C_{is} \geq 0$$

It may be verified that $C_{is} = X^{*}_{is}$ solves this problem,[5] thus the prices $\{P^{*}_{sc}\}$ may be viewed as the *implicit* equilibrium spot prices of the first exchange regime.

However, if each consumer i knew prior to the realization of the state that he would have the opportunity to trade in spot markets at predetermined prices $\{P_{sc}\}$ as well as in markets for contingent claims at prices $\{r_{sc}\}$, each would solve the following recursive problem. First given income Y_{is} in state s in terms of the numeraire, commodity C, each consumer i would maximize $U^i(C_{is})$ with respect to C_{is} subject to the budget constraint $P_s \cdot C_{is} \leq Y_{is}$ with $C_{is} \geq 0$. Let $h_{is}(Y_{is}, P_s)$ denote the maximizing choice of C_{is}. Then define the indirect utility function $V^i(Y_{is}, P_s) = U^i[h_{is}(Y_{is}, P_s)]$. But Y_{is} is determined by the claims $\{X_{isc}\}$ acquired in the prestate markets for contingent claims. That is,

$$(5) \qquad Y_{is} = \sum_{c=1}^{C} P_{sc} X_{isc}$$

Hence in the market for contingent claims consumer i would maximize

$$(6) \qquad \sum_{s=1}^{S} \pi_s V^i\left(\sum_{c=1}^{C} P_{sc} X_{isc}, P_s\right)$$

with respect to $\{X_{isc}\}$ subject to the budget constraint (1) and income constraints $Y_{is} \geq 0$.[6] It should be clear that a maximizing choice $\{X^{*}_{isc}\}$ for this recursive problem at prices $\{P_{sc}\}$ and $\{r_{sc}\}$ cannot be unique. For if $\{X^{*}_{isc}\}$ were a maximizing choice, so also would be all bundles $\{X^{**}_{isc}\}$ such that

$$\sum_{c=1}^{C} P_{sc} X^{**}_{isc} = \sum_{c=1}^{C} P_{sc} X^{*}_{isc}$$

for each state s. Roughly speaking, given the opportunity to trade in spot markets at spot prices $\{P_{sc}\}$, consumer i cares only about the income he will have in the various states.

The indeterminacy of the recursive problem just described suggests that some further restrictions can be placed on trades without altering the ability of the consumer to acquire (ultimately) the maximizing consumption bundles. Indeed one such restriction was placed on the consumer in the first exchange regime—that there be no trading in spot markets.[7] This paper examines restrictions associated with forward contracts.

[5] For suppose $C_{is} = X^{**}_{is}$ solves this problem with $U^i(X^{**}_{is}) > U^i(X^{*}_{is})$ and $\sum_{c=1}^{C} (X^{**}_{isc} - X^{*}_{isc}) P^{*}_{sc} = 0$. Then from (4) one obtains $\sum_{c=1}^{C} (X^{**}_{isc} - X^{*}_{isc}) r^{*}_{sc} = 0$. This in conjunction with (1) establishes that $\{X^{**}_{isc}\}$ was obtainable in the first regime but not chosen, the desired contradiction.

[6] Here and below, these income constraints rule out bankruptcy; each consumer is assumed to honor all contracts into which he has entered, and with these constraints each has sufficient income to do so. However, it is *not* required that delivery be made in spot markets of commodities sold in the markets for claims; it is supposed that each consumer accepts delivery of all commodity bundles which when valued at spot prices yield incomes equivalent to the yield of the claim in question. It can also be established that under previous assumptions $V^i(\cdot, P_s)$ is strictly concave and continuously differentiable with $V^i_1(0, P_s) = \infty$. Hence in a maximizing position $Y_{is} > 0$ and the income constraints need not be made explicit.

[7] It can be established rigorously that such a restriction is not constraining.

For each consumer i and each commodity c these restrictions are of the form $(X_{isc} - Z_{isc}) = (X_{itc} - Z_{itc})$ for all states s and t. Thus for example if consumer i purchases a specified number of claims on commodity c contingent on state s, then he must also purchase the same number of claims on commodity c contingent on all other states. In effect only unconditional claims can be purchased or issued in such forward markets.

Thus in the second exchange regime of this paper each consumer can trade unconditional forward contracts in prestate markets and can also trade in poststate spot markets. The decision problem which confronts a consumer in such a regime is now formalized. Let Q_{ic} denote the number of unconditional claims on commodity c purchased forward by consumer i in forward markets. (Thus if Q_{ic} is negative, commodity c is sold forward.) Let f_c denote the forward price of an unconditional unit claim on commodity c in terms of some abstract unit of account. Then the budget constraint for consumer i in forward markets is

$$\sum_{c=1}^{C} f_c Q_{ic} = 0 \tag{7}$$

Having acquired forward contracts $\{Q_{ic}\}$, consumer i enters spot market s with income

$$Y_{is} = \sum_{c=1}^{C} P_{sc} Z_{isc} + \sum_{c=1}^{C} P_{sc} Q_{ic} \tag{8}$$

Thus, with trading permitted in spot markets, consumer i maximizes

$$\sum_{s=1}^{S} \pi_s V^i\left(\sum_{c=1}^{C} P_{sc}(Z_{isc} + Q_{ic}), P_s\right) \tag{9}$$

with respect to $\{Q_{ic}\}$ subject to the budget constraint (7).

An equilibrium of the second exchange regime is a set of forward prices $\{f_c^*\}$, a set of spot prices $\{P_{sc}^*\}$, a forward position $\{Q_{ic}^*\}$ $i = 1,2,\ldots,I$ and a consumption allocation $\{X_{isc}^*\}$ $i = 1,2,\ldots,I$, such that $\{Q_{ic}^*\}$ and $\{X_{isc}^*\}$ are maximizing for each consumer i in forward markets and spot markets, respectively. That is $\{Q_{ic}^*\}$ maximizes (9) subject to (7) under $\{f_c^*\}$ $\{P_{sc}^*\}$ with

$$X_{is}^* = h_{is}\left(\sum_{c=1}^{C} P_{sc}^*(Z_{isc} + Q_{ic}^*), P_s^*\right)$$

Forward markets clear for each commodity c,

$$\sum_{i=1}^{I} Q_{ic}^* = 0 \qquad c = 1,2,\ldots,C \tag{10}$$

and spot markets clear for each commodity c in each state s,

$$\sum_{i=1}^{I} [X_{isc}^* - (Z_{isc} + Q_{ic}^*)] = 0 \tag{11}$$
$$s = 1,2,\ldots,S \qquad c = 1,2,\ldots,C$$

II. On the Equivalence of the Two Exchange Regimes

In this section it will be argued that, subject to some restrictions on the matrix of (implicit) spot prices, any Pareto optimal allocation can be supported as a competitive equilibrium of the second exchange regime with suitable redistribution of endowments and that competitive equilibria of the second regime are Pareto optimal. More formally we have

PROPOSITION 1: *Suppose that a Pareto optimal allocation* $\{X_{isc}^*\}$ $i = 1,2,\ldots,I$ *can be supported as a competitive equilibrium of the first exchange regime with endowments* $\{Z_{isc}\}$ $i = 1,2,\ldots,I$ *and claim prices* $\{r_{sc}^*\}$ *such that the* $S \times C$ *matrix* $p'' = [P_{sc}^*]$ *(where* $P_{sc}^* = r_{sc}^*/r_{sC}^*$*) is of rank* S*. Then* $\{X_{isc}^*\}$ $i = 1, 2,\ldots I$ *can be supported with the same endowments as a competitive equilibrium of the second exchange regime with forward markets in S commodities.*

PROPOSITION 2: *Suppose there exists a competitive equilibrium of the second exchange regime with forward markets in S commodities with spot prices* $\{P_{sc}^*\}$ *and consumption allocation* $\{X_{isc}^*\}$ $i = 1,2,\ldots,I$ *such that the corresponding* $S \times S$ *matrix* $P = [P_{sc}^*]$ *is of rank* S*. Then the consumption allocation* $\{X_{isc}^*\}$ $i = 1,2,\ldots,I$ *is Pareto optimal.*

The formal proofs of these propositions are contained in Appendix A, but are now outlined with some motivating remarks. As for the first proposition it is clear from the classical theorems of welfare economics that any Pareto optimal allocation can be supported as a competitive equilibrium of the first exchange regime with suitable redistribution of endowments in the various states. Having specified this same distribution of endowments, the second exchange regime is imposed. Then it is argued that each consumer is endowed implicitly with forward contracts; each consumer can issue forward contracts up to his ability to honor such claims with his income in the various states. Also, each consumer must have sufficient income in the various states to purchase the optimal allocation assigned to him. This determines the forward contracts he must acquire in the forward markets. It is then shown that at appropriately selected spot and forward prices the resulting excess demands are consistent with the budget constraint of each consumer and that the acquired forward contracts are indeed maximizing. Finally it is shown that all markets clear.

Proposition 1 also has an important corollary:

If there exists a competitive equilibrium of the first exchange regime such that the matrix P″ is of rank S, then there exists a competitive equilibrium of the second exchange regime with a consumption allocation which is Pareto optimal. This follows from the fact that the equilibrium allocation of the first regime is Pareto optimal and by hypothesis can be supported in a competitive equilibrium without any redistribution of endowments.

The idea underlying the proof of the second proposition is that an equilibrium consumption allocation of the second regime can be supported as an equilibrium allocation of the first regime and hence is Pareto optimal.

It remains to examine the hypothesis that the matrix of spot prices have rank equal to the number of states. Essentially this hypothesis ensures that the returns of forward contracts span the space of all possible returns. In order to clarify the role of this spanning property two examples are now described, one with the spanning property and one without.

For the first example there are three commodities and three states and the 3 x 3 matrix of spot prices is assumed to be of rank three. Suppose that a nonnegative vector of incomes $\{Y_{is}; s = 1,2,3\}$ is to be attained by a forward position $\{Q_{ic}; c = 1,2,3\}$. Then equations (7) and (8) are of the form

$$\sum_{c=1}^{3} f_c Q_{ic} = 0 \tag{12}$$

$$Y_{is} = \sum_{c=1}^{3} P_{sc} Z_{isc} + \sum_{c=1}^{3} P_{sc} Q_{ic} \qquad s = 1,2,3 \tag{13}$$

Setting $f_3 = 1$, solving for Q_{i3} in (12), and substituting into (13) yields

$$Y_{is} = \hat{Y}_{is} + \sum_{c=1}^{2} (P_{sc} - f_c) Q_{ic} \qquad s = 1,2,3 \tag{14}$$

$$Y_{is} = \sum_{c=1}^{3} P_{sc} Z_{isc} \qquad s = 1,2,3 \tag{15}$$

Equation (14) is a parametric representation of a plane in three space through the endowed state distribution income point $\{\hat{Y}_{is}\}$. With suitable specification of the spot and forward prices, each consumer i can exchange income in any one state for income in any other without altering income in the third as illustrated in Figure 1. Thus in effect in the second regime each consumer i maximizes

$$\sum_{s=1}^{3} \pi_s V^i(Y_{is}, P_s)$$

with respect to $\{Y_{is}\}$ as determined by the choice of $\{Q_{ic}\}$ subject to constraints (14). Moreover as each consumer i is confronted with a budgetplane with the same gradient (determined by the prices $\{f_c\}$ and $\{P_{sc}\}$) each will have the same rate of substitution

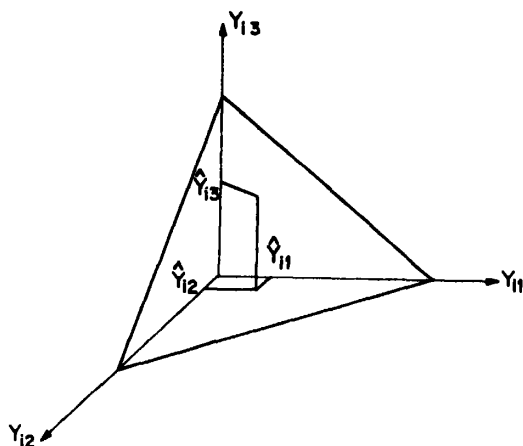



FIGURE 1

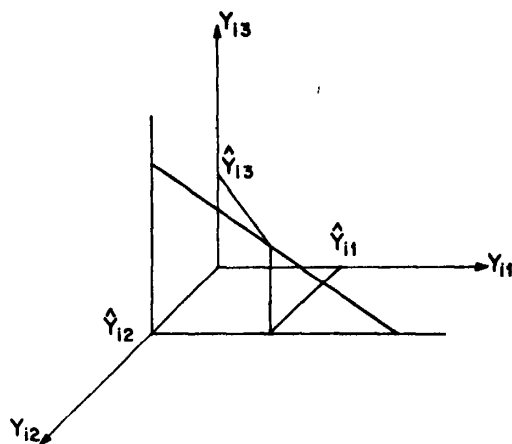



FIGURE 2

of income (the numeraire good) across states in an equilibrium, and the equilibrium allocation will be optimal.

For the second example there are three states but only two goods so that the matrix of spot prices cannot be of rank three. Then setting $f_2 = 1$, the analogue of (13) is of the form

$$Y_{is} = \hat{Y}_{is} + Q_{i1}(P_{s1} - f_1) \qquad s = 1,2,3 \tag{16}$$

Equation (16) is a parametric representation of a line in three space. If for example $P_{11} < f_1 = P_{21} < P_{31}$ it is impossible to alter income in the second state. This is illustrated in Figure 2. If for example $P_{11} < f_1 < P_{21} < P_{31}$, it is impossible to alter income in one state without altering income in the other two. Though in an equilibrium of the second regime each consumer is confronted with a budget line of the same slope (determined by the prices $\{f_c\}$ and $\{P_{sc}\}$), in general each will not have the same rate of substitution of income across states. This example thus illustrates the potential for inefficiency when states outnumber commodities.

An attempt is now made to relate Proposition 1 to the results of Arrow. His principal conclusion is that an optimal allocation risk bearing can be achieved in a distribution economy (with money) by competitive markets in elementary securities. He emphasizes that a security is a claim payable in money in contrast to claims against specific commodities. But of course in the context of an exchange economy money can be no more than a numeraire. Thus, for example, if the Cth good is selected as the numeraire of each spot market, an elementary security yielding one monetary unit in state s and zero otherwise can be nothing other than a claim on commodity C in state s. It is shown in Appendix B of this paper that in an exchange economy any optimal allocation can be achieved with a set of S securities with linearly independent returns, where these returns are in terms of the amount of the numeraire good which a bearer can purchase in the spot market of each state. This then is the generalized analogue of Arrow's theorem for an exchange economy. Arrow-Debreu securities (claims on the numeraire good only) can be viewed as a particularly simple set of such securities. And subject to a rank condition on a matrix of spot prices P, forward contracts also constitute a spanning set. A forward purchase on commodity c for example has state dependent return represented by the cth column of the matrix P. The condition that P be of rank S is equivalent to the condition that the column vectors of returns of the S forward "securities" be linearly independent.

III. Forward Trading as General Equilibrium Hedging

This section is intended to give some further insight into the workings of forward markets in a general equilibrium setting. A simple example is presented which illustrates that with active spot markets, forward contracts serve as a hedge against exogenously random endowments *and* endogenously random spot prices. This general equilibrium hedging model of forward markets may be contrasted with the classic partial equilibrium approach of John M. Keynes and John Hicks which emphasizes a distinction between hedgers and speculators. In particular in the model of this paper maximizing behavior on the part of risk-averse agents does not necessarily involve the elimination of risk by purchasing the consumption bundle forward. The example also allows some inferences concerning the existence and optimality of a competitive equilibrium of the second exchange regime when market structure is incomplete.

For the example there are two representative consumers, S states of the world ($S \geq 2$), and two commodities. The first consumer is endowed with the first commodity only, and the second consumer is endowed with the second commodity only. That is, $Z_{isc} = 0$ if $i \neq c$. Without loss of generality it is supposed that Z_{1s1} is strictly increasing in s. It is also assumed that Z_{2s2} is equal to some constant Z_2 for all states.

Preferences are identical for both consumers. Each has a utility function of the form $(U(\ ,\ ) = g[W(\ ,\ )]$ where $W(\ ,\ )$ displays constant elasticity of substitution and $g(\ )$ is a monotone increasing function. Hence $W(\ ,\ )$ is of the form

$$W(C_{is1}, C_{is2}) = [(\alpha)C_{is1}^{-\rho} + (1 - \alpha)C_{is2}^{-\rho}]^{-1/\rho} \quad \text{if } \sigma \neq 1$$

$$W(C_{is1}, C_{is2}) = C_{is1}^{\alpha} C_{is2}^{1-\alpha} \quad \text{if } \sigma = 1$$

where σ, the elasticity of substitution, equals $1/(1 + \rho)$ and $0 < \alpha < 1$. It is further assumed that $g(W) = W^{\mu}$ where $0 < \mu < 1$, or $g(W) = ln\, W$.

Let the second commodity be chosen as the numeraire in the forward markets. Then each consumer i maximizes

$$(17)\quad \sum_{s=1}^{S} \pi_s V\left(\sum_{c=1}^{2} Z_{isc}P_{sc} + Q_{i1}(P_{s1} - f_1), P_s\right)$$

with respect to Q_{i1}, yielding necessary and sufficient first-order conditions

$$(18)\quad \sum_{s=1}^{S} \pi_s(P_{s1} - f_1)V_1\left\{\sum_{c=1}^{2} Z_{isc}P_{sc} + \psi^i(f_1)[P_{s1} - f_1], P_s\right\} = 0$$

where $\psi^i(f_1)$ denotes the maximizing choice of Q_{i1} as a function of f_1. It can be shown that $\psi_1^i(f_1) < 0$.[8]

It also can be shown that for this example equilibrium spot prices are independent of the existence and direction of forward trading as

$$(19)\quad P_{s1} = \left(\frac{\alpha}{1 - \alpha}\right)\left(\frac{Z_2}{Z_{1s1}}\right)^{1/\sigma}$$

Consequently P_{s1} is strictly decreasing in s. It also follows that

$$(20)\quad P_{s1}Z_{1s1} = \left(\frac{\alpha}{1 - \alpha}\right) Z_2^{1/\sigma} Z_{1s1}^{(\sigma-1)/\sigma}$$

Equation (20), which displays the value of the exogenous endowment of the first consumer as a function of s and σ, will be useful in what follows.

There remains the task of establishing the existence and direction of equilibrium forward trading. We have

PROPOSITION 3: *Under the assumptions of the example there exists a competitive*

[8]The objective function is strictly concave and continuously differentiable in Q_{i1}. Moreover the income constraints $Y_{is} \geq 0$ restrict the choice of Q_{i1} to a compact set. Hence there exists a unique maximizing choice of Q_{i1}. Also with $V_1^i(0, P_s) = \infty$, this choice must be an interior solution and the implicit function theorem applies. With decreasing absolute risk aversion, the derivative of $\psi^i(\ )$ can be signed.

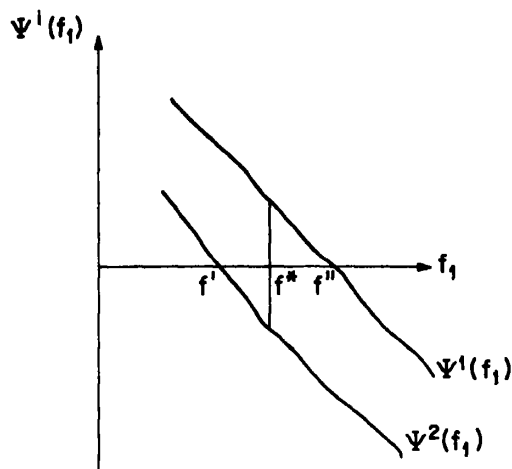



FIGURE 3

equilibrium of the second regime. Moreover

(i) *if* $\sigma > 1$, *then* $Q_{11} > 0$

(ii) *if* $0 < \sigma < 1$, *then* $Q_{11} < 0$

(iii) *if* $\sigma = 1$, *then* $Q_{11} = 0$

The idea underlying the proof of the proposition is illustrated in Figure 3. The object is to find forward prices f'' and f' at which the first and second consumer, respectively, would not wish to trade, and to show these prices differ in an appropriate way. The equilibrium price f^* can then be found and the properties of the proposition verified. A formal proof of the proposition is contained in Appendix C.

The results of the proposition are not as counterintuitive as they may first seem. Consider the case $\sigma > 1$. From (20), $P_{s1}Z_{1s1}$ is strictly increasing in s. Hence the first consumer is relatively more anxious to engage in a venture which is strictly decreasing in s than is the second consumer. Forward purchases of the first commodity with per unit return $(P_{s1} - f_1)$ represents such a venture. Thus each consumer purchases forward the single commodity with which he is endowed. Maximizing behavior in this general equilibrium hedging model need not entail purchasing the consumption bundle forward.

What can be said of the optimality of a competitive equilibrium allocation of the second exchange regime when market structure may be incomplete (as in the example of this section with $S > 2$)? If tastes are identical and homothetic (as in the example), it is possible to make some welfare comparisons. For if tastes are identical and homothetic, spot market prices are independent of the existence and direction of prestate forward trading. If there are forward markets and if a consumer chooses not to participate in such markets, then his consumption possibility set is precisely what it would have been had there been no forward markets at all. Hence the possibility of forward trading can only make him better off. This yields:

PROPOSITION 4: *If tastes are identical and homothetic, then a competitive equilibrium allocation of the second exchange regime is Pareto noninferior and possibly Pareto superior to the competitive equilibrium allocation with all markets for claims prohibited.*

IV. Concluding Remarks

Jacques Drèze has stressed the need for research into the functions and shortcomings of existing institutions and for the application of standard welfare economics based on Pareto optimality to limited exchange opportunities for risk bearing. The objective of this paper was to examine the workings and welfare implications of forward markets and to place those markets in the context of complete markets for contingent claims. It was found that with at least as many commodities as states, prestate forward markets with poststate spot markets may support Pareto optimal allocations. Thus the existence of forward markets in some commodities rather than markets for contingent claims should not be taken as prima facie evidence of some inefficiency.[9]

[9]The ultimate intent of a paper of this sort is to explain why futures contracts with subsequent spot markets is a prominent institutional configuration. If agents were indifferent between complete markets for contingent claims and futures contracts with subse-

APPENDIX A

PROOF of Proposition 1:

As P'' is of rank S, C-S columns may be deleted from P'' while leaving a square matrix P of rank S. Then without loss of generality commodities with prices in P are numbered one through S. Let $\hat{Y}_{is} = \sum_{c=1}^{C} P^*_{sc} Z_{isc}$ with associated S x 1 vector $\hat{Y}_i$. Then each consumer i is endowed implicitly with forward contracts $\{E_{ic};\ c = 1, 2, \ldots, S\}$ with associated S x 1 vector E_i such that

$$\text{(A1)}\quad \sum_{c=1}^{S} P^*_{sc} E_{ic} = \hat{Y}_{is} \qquad s = 1, 2, \ldots, S$$

or in matrix notation $PE_i = \hat{Y}_i$. Let $Y^*_{is} = \sum_{c=1}^{C} P^*_{sc} X^*_{isc}$ with associated S x 1 vector Y^*_i. Then if the optimal allocation $\{X^*_{isc}\} i = 1, 2, \ldots, I$ is to be achieved consumer i must enter spot markets holding forward contracts $\{F_{ic};\ c = 1, 2, \ldots, S\}$ with associated S x 1 vector F_i such that

$$\text{(A2)}\quad \sum_{c=1}^{S} P^*_{sc} F_{ic} = Y^*_{is} \qquad s = 1, 2, \ldots, S$$

or in matrix notation $PF_i = Y^*_i$.

Choose spot prices $P_{sc} = P^*_{sc}$ and choose forward prices $f_c = f^*_c$, $c = 1, 2, \ldots S$, where $f^*_c = \sum_{s=1}^{S} r^*_{sc}$. Define a S x S diagonal matrix D with diagonal elements r^*_{sc} and zeros elsewhere.

First it is shown that individual budget constraints are satisfied. From (A1) and (A2) $DPQ^*_i = D(Y^*_i - \hat{Y}_i)$ where $Q^*_i = F_i - E_i$, with typical row s,

$$\text{(A3)}\quad \sum_{c=1}^{S} r^*_{sc} Q^*_{ic} = \sum_{c=1}^{C} r^*_{sc}(X^*_{isc} - Z_{isc})$$

Summing over the rows (A3) yields

$$\text{(A4)}\quad \sum_{c=1}^{S} f^*_c Q^*_{ic} = \sum_{s=1}^{S}\sum_{c=1}^{C} r^*_{sc}(X^*_{isc} - Z_{isc})$$

By hypothesis the right side of (A4) equals zero, and hence so does the left side.

Now suppose that Q^*_i were not a maximizing forward position for some consumer i given prices $\{f^*_c\}$, $\{P^*_{sc}\}$. That is, suppose there existed some choice Q^{**}_i of forward contracts and associated consumption $\{X^{**}_{isc}\}$ in spot markets such that

$$\text{(A5)}\quad \sum_{s=1}^{S} \pi_s U^i(X^{**}_{is}) > \sum_{s=1}^{S} \pi_s U^i(X^*_{is})$$

Since these choices are feasible, the budget constraint in forward markets is satisfied, i.e., $\sum_{c=1}^{S} f^*_c Q^{**}_{ic} = 0$, and there is sufficient income to purchase $\{X^{**}_{isc}\}$ in spot markets, i.e., $PQ^{**}_i = Y^{**}_i - \hat{Y}_i$ where Y^{**}_i is the S x 1 vector associated with $Y^{**}_{is} = \sum_{c=1}^{C} P_{sc} X^{**}_{isc}$. With virtually the same manipulations that yielded (A4), one obtains

$$\text{(A6)}\quad \sum_{c=1}^{S} f^*_c Q^{**}_{ic} = \sum_{s=1}^{S}\sum_{c=1}^{C} r^*_{sc}(X^{**}_{isc} - Z_{isc})$$

But

$$\text{(A7)}\quad \sum_{c=1}^{S} f^*_c Q^{**}_{ic} = 0$$

and therefore

$$\text{(A8)}\quad \sum_{s=1}^{S}\sum_{c=1}^{C} r^*_{sc}(X^{**}_{isc} - Z_{isc}) = 0$$

so that $\{X^{**}_{isc}\}$ was feasible under the budget constraint of the first regime. This is the desired contradiction.

It remains to show that forward markets clear. From (A1) and (A2)

quent spot markets, and if there were a cost associated with the former contracts which is not associated with the latter, then one structure would emerge endogenously. A cost which might be associated with contingent but not with futures contracts could be the cost of state verification. It is in this sense that the requirement that P'' be of rank S is somewhat disappointing. If P'' is of rank S, then no two rows of P'' can be identical. Agents will be fully informed by the spot market prices of which state has occurred. State verification is costless and is no obstacle to the making of contingent contracts. Futures contracts with subsequent spot markets may allow agents to do just as well, but there is nothing in the model to lead them to choose one structure over the other.

$$\text{(A9)} \quad \sum_{i=1}^{I} Q_i^* = P^{-1} \sum_{i=1}^{I} (Y_i^* - \hat{Y}_i)$$

But

(A10)

$$\sum_{i=1}^{I} (Y_{is}^* - \hat{Y}_{is}) = \sum_{c=1}^{C} P_{sc}^* \sum_{i=1}^{I} (X_{isc}^* - Z_{isc})$$

From (3) the right side equals zero. Substitution into (A9) yields the desired result.

Finally, each spot market s is in equilibrium at the prices $\{P_{sc}^*\}$. For at these prices consumers achieve the same distribution of incomes across states as in the equilibrium of the first regime. That is, each consumer is on the same budget hyperplane in each state. As spot markets were implicitly in equilibrium at these same prices (see Section I) they will continue to be so (see (A15) below).

PROOF of Proposition 2:

The idea underlying the proof is that the consumption allocation of the second regime can be obtained as an equilibrium allocation of the first regime. Without loss of generality assume the first S commodities are traded forward in the second regime. Let $\{f_c^*\}$ and $\{Q_{ic}^*\}$ $c = 1,2,\ldots,S$; $i = 1,2,\ldots,I$ denote the equilibrium forward prices and forward positions, respectively, of the second regime. Then let the claim prices $\{r_{sc}^*\}$ be chosen such that

$$\text{(A11)} \quad f_c^* = \sum_{s=1}^{S} r_{sc}^* \qquad c = 1,2,\ldots,S$$

$$\text{(A12)} \quad r_{sc}^* = P_{sc}^* r_{sC}^* \qquad s = 1,2,\ldots,S; \quad c = 1,2,\ldots,C$$

(Note that by substituting (A12) into (A11) one obtains the system

$$\text{(A13)} \quad f_c^* = \sum_{s=1}^{S} P_{sc}^* r_{sC}^* \qquad c = 1,2,\ldots,S$$

With P of full rank, there exist a unique solution for $\{r_{sC}^* s = 1,2,\ldots,S\}$ in (A13) so (A11) and (A12) are well defined.)

As $\{X_{isc}^*\}$ is the final allocation, it must be, as in the proof of Proposition 1, that (A4) holds. But by hypothesis the left side of (A4) equals zero. Hence $\{X_{isc}^*\}$ satisfies the budget constraint (1) under $\{r_{sc}^*\}$.

Now suppose $\{X_{isc}^*\}$ were not maximizing under the first regime. Suppose there exist some $\{X_{isc}^{**}\}$ such that (A8) and (A5) hold. But then define $\{Q_{ic}^{**}\}$ such that there is sufficient income to purchase $\{X_{isc}^{**}\}$. That is, (A6) applies. From (A8), the right side of (A6) equals zero. Hence $\sum_{c=1}^{S} f_c^* Q_{ic}^{**} = 0$. This contradicts $\{Q_{ic}^*\}$ as maximizing.

Finally note that the markets for claims are in equilibrium. For let $\{\hat{X}_{isc}\}$ denote the forward position of consumer i in the second regime after trading in forward markets but before trading in spot markets. Then

$$\text{(A14)} \quad X_{isc}^* = (Z_{isc} + Q_{ic}^*) + (X_{isc}^* - \hat{X}_{isc}) \qquad s = 1,2,\ldots,S; \quad c = 1,2,\ldots,C$$

Summing (A14) over i yields

$$\text{(A15)} \quad \sum_{i=1}^{I} (X_{isc}^* - Z_{isc}) = \sum_{i=1}^{I} Q_{ic}^* + \sum_{i=1}^{I} (X_{isc}^* - \hat{X}_{isc}) \qquad s = 1,2,\ldots,S; \quad c = 1,2,\ldots,C$$

As forward and spot markets clear in the second regime, the right side of (A15) equals zero.

B

In what follows a security is defined to be a linear combination of unit claims on the SC contingent commodities. That is, a security of type τ entitles the holder to β_{sc}^{τ} units of commodity c in state s, $c = 1,2,\ldots,C$, $s = 1,2,\ldots,S$. Let $R_{s\tau}$ denote the return (in terms of commodity C) of security τ in state s so that given spot prices $\{P_{sc}\}$, $R_{s\tau} = \sum_{c=1}^{C} P_{sc}\beta_{sc}^{\tau}$.

PROPOSITION B-1: *Suppose that a Pareto optimal allocation* $\{X_{isc}^*\}$ $i = 1,2,\ldots,I$ *can be supported as a competitive equilibrium with complete markets for contingent claims and with no trade in spot markets with en-*

downments $\{Z_{isc}\}$ $i = 1,2,\ldots,I$, and claim prices $\{r^*_{sc}\}$. Suppose also that there exist S securities where security of type τ has return $R^*_{s\tau}$ in state s as determined by the spot prices $P^*_{sc} = r^*_{sc}/r^*_{sC}$ such that the $S \times S$ matrix of security returns $R = [R^*_{s\tau}]$ is of rank S. Then $\{X^*_{isc}\}$ $i = 1,2,\ldots,I$ can be supported with the same endowments as a competitive equilibrium with prestate markets for the S securities and poststate spot markets.

PROOF:

Let $P_{sc} = P^*_{sc}$. Let $\hat{Y}_{is} = \sum_{c=1}^{C} P^*_{sc} Z_{isc}$ with associated $S \times 1$ vector $\hat{Y}_i$. Then each consumer i is endowed implicitly with $\hat{E}_{i\tau}$ units of security of type τ with associated $S \times 1$ vector $\hat{E}_i$ defined by

$$(A16) \qquad R\hat{E}_i = \hat{Y}_i$$

Let $Y^*_{is} = \sum_{c=1}^{C} P^*_{sc} X^*_{isc}$ with associated $S \times 1$ vector Y^*_i. Then if the allocation $\{X^*_{isc}\}$ is to be attained, consumer i must enter spot markets with securities $\tilde{F}_{i\tau}$ with associated $S \times 1$ vector $\tilde{F}_i$ defined by

$$(A17) \qquad R\tilde{F}_i = Y^*_i$$

Subtracting (A17) from (A16), premultiplying by the $S \times S$ diagonal matrix D with elements r^*_{sC} and summing over rows yields

$$(A18) \qquad \sum_{\tau=1}^{S} \tilde{f}^*_{\tau}(\tilde{F}_{i\tau} - \hat{E}_{i\tau}) = \sum_{s=1}^{S}\sum_{c=1}^{C} r^*_{sc}(X^*_{isc} - Z_{isc})$$

where $\tilde{f}^*_{\tau} = \sum_{s=1}^{S} r^*_{sC} R^*_{s\tau}$ is taken as the price of security τ. By hypothesis the right side of (A18) equals zero so the $\tilde{Q}^*_i = \tilde{F}_i - \hat{E}_i$ security trades are consistent with the budget constraint of consumer i in the prestate security markets.

Moreover $\{\tilde{Q}^*_i\}$ is maximizing for consumer i. The argument is virtually identical to the one given in Proposition 1 with E_i, F_i, Q^*_i, f^*_c, and P replaced by $\hat{E}_i$, $\tilde{F}_i$, $\tilde{Q}^*_i$, $\tilde{f}^*_{\tau}$, and R, respectively.

Also, security markets clear. The excess demand for security τ is

$$(A19) \qquad \sum_{i=1}^{I} (\tilde{F}_{i\tau} - \hat{E}_{i\tau})$$

From (A16), $\hat{E}_i = R^{-1}\hat{Y}_i$, using the fact that R is of full rank. That is,

$$(A20) \qquad \hat{E}_{i\tau} = \sum_{s=1}^{S} \alpha_{\tau s}\hat{Y}_{is}$$

where the $\{\alpha_{\tau s}\}$ are expressions involving the terms of R. These may be regarded as constants. Similarly one obtains

$$(A21) \qquad \tilde{F}_{i\tau} = \sum_{s=1}^{S} \alpha_{\tau s} Y^*_{is}$$

Then substituting (A20) and (A21) into (A19) and recalling the definitions of Y^*_{is} and $\hat{Y}_{is}$ one obtains

$$(A22) \qquad \sum_{s=1}^{S} \alpha_{\tau s} \sum_{c=1}^{C} P^*_{sc} \sum_{i=1}^{I} (X^*_{isc} - Z_{isc})$$

which equals zero by the market-clearing conditions of the first regime.

Finally it may be argued as in Proposition 1 that spot markets clear.

PROPOSITION B-2: *Suppose that there exists a competitive equilibrium with prestate markets in S securities and with poststate spot markets with spot prices $\{P^*_{sc}\}$ and a consumption allocation $\{X^*_{isc}\}$ $i = 1,2,\ldots,I$ such that the matrix of security returns $R = [R^*_{s\tau}]$ is of rank S. Then the consumption allocation $\{X^*_{isc}\}$ $i = 1,2,\ldots,I$ is Pareto optimal.*

PROOF:

Let $\{\tilde{f}^*_{\tau}\}$ and $\{\tilde{Q}^*_{i\tau}\}$ $i = 1,2,\ldots,I$ denote the equilibrium security prices and security trades, respectively. Then choose claim prices $\{r^*_{sc}\}$ to satisfy

$$(A23) \qquad \tilde{f}^*_{\tau} = \sum_{s=1}^{S} r^*_{sC}R^*_{s\tau} \qquad \tau = 1,2,\ldots,S$$

$$(A24) \qquad r^*_{sc} = P^*_{sc} r^*_{sC}$$

(As R is of full rank, these equations are well defined.) Then as in Proposition 2, it can be shown that $\{X^*_{isc}\}$ is a maximizing choice of each consumer i in claims markets of the first regime. Finally let $\{\bar{X}_{isc}\}$ denote the implicit forward position of consumer i after trading in security markets. Then

(A25)
$$X^*_{isc} = Z_{isc} + \sum_{\tau=1}^{S} \beta^{\tau}_{sc} \tilde{Q}^*_{i\tau} + (X^*_{isc} - \hat{X}_{isc})$$
$$s = 1,2,\ldots,S; \qquad c = 1,2,\ldots,C$$

Summing over i in (A25) yields

(A26)
$$\sum_{i=1}^{I} (X^*_{isc} - Z_{isc}) = \sum_{\tau=1}^{S} \beta^{\tau}_{sc} \sum_{i=1}^{I} \tilde{Q}^*_{i\tau} + \sum_{i=1}^{I} (X^*_{isc} - \hat{X}_{isc})$$
$$s = 1,2,\ldots,S; \qquad c = 1,2,\ldots,C$$

As security markets and spot markets clear, the right side of (A26) equals zero.

Two special cases of the propositions should be noted. First if there exist S commodities such that the corresponding $S \times S$ matrix P of spot prices is of full rank, setting $R = P$, Propositions 1 and 2 follow. Also with S elementary Arrow-Debreu securities (where a security of type s yields one unit of commodity C in state s and zero otherwise) $R = I$, the identity matrix, and the propositions apply.

C

PROOF of Proposition 3:

Under the assumptions of the example the indirect utility function is of one of the following two forms:

(A27) $$V(Y_{is}, P_s) = \varphi(P_s)^{\mu}(Y_{is})^{\mu}$$

(A28) $$V(Y_{is}, P_s) = \ln Y_{is} + \ln \varphi(P_s)$$

where $\varphi(P_s)$ is an expression in terms of α, σ, and P_s. Define

(A29)
$$G^i(Q_{i1}, f_1) = \sum_{s=1}^{S} \pi_s(P_{s1} - f_1) V_1\left(\sum_{c=1}^{2} P_{sc}Z_{isc} + Q_{i1}(P_{s1} - f_1), P_s\right)$$

Let f'' be defined by the equation $G^1(0, f'') = 0$ so that

(A30)
$$\sum_{s=1}^{S} \pi_s(P_{s1} - f'')\mu(P_{s1}Z_{1s1})^{\mu-1}\varphi(P_s)^{\mu} = 0$$

(A31) $$\sum_{s=1}^{S} [\pi_s(P_{s1} - f'')]/[P_{s1}Z_{1s1}] = 0$$

for forms (A27) and (A28) of $V(\cdot,\cdot)$, respectively. Let f' be defined by the equation $G^2(0, f') = 0$ so that

(A32) $$\sum_{s=1}^{S} \pi_s(P_{s1} - f')\mu(Z_2)^{\mu-1}\varphi(P_s)^{\mu} = 0$$

(A33) $$\sum_{s=1}^{S} [\pi_s(P_{s1} - f')]/[Z_2] = 0$$

for forms (A27) and (A28) of $V(\cdot,\cdot)$, respectively. Now consider the following cases:

CASE (i): $\sigma > 1$

With $\sigma > 1$ it follows from (20) that $P_{s1}Z_{1s1}$ is strictly increasing in s. Therefore, both forms (A27) and (A28), $f'' > f'$. Let $\psi(f_1) = \sum_{i=1}^{2} \psi^i(f_1)$. Then $\psi(f_1)$ is continuous. As $d\psi^i(f_1)/df_1 < 0$, $i = 1, 2$, $\psi(f'') < 0$ and $\psi(f') > 0$; see Figure 3. Therefore, there exists some f^*, $f' < f^* < f''$, with $\psi(f^*) = 0$. Hence f^* is the unique equilibrium forward price with $\psi^1(f^*) > 0$.

CASE (ii): $0 < \sigma < 1$

With $0 < \sigma < 1$, $P_{s1}Z_{1s1}$ is strictly decreasing in s. Consequently $f'' < f'$ and there exist some f^*, $f'' < f^* < f'$, with $\psi^1(f^*) < 0$.

CASE (iii): $\sigma = 1$

With $\sigma = 1$, $P_{s1}Z_{1s1}$ is constant in s so that $f^* = f' = f''$ and $\psi^i(f^*) = 0$.

REFERENCES

K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev. Econ. Stud.*, Apr. 1964, *31*, 91–96.

J. Drèze, "Econometrics and Decision Theory," *Econometrica*, Jan. 1972, *40*, 1–17.

S. Ekern and R. Wilson, "On the Theory of the

Firm in an Economy with Incomplete Markets," *Bell J. Econ.*, Spring 1974, *5*, 171–80.

John R. Hicks, *Value and Capital*, Oxford 1946, 135–39.

John M. Keynes, *A Treatise on Money*, London 1950, 142–47.

R. Radner, "A Note on Unanimity of Stockholders' Preferences among Alternative Production Plans: A Reformulation of the Ekern-Wilson Model," *Bell J. Econ.*, Spring 1974, *5*, 181–84.

S. A. Ross, "Options and Efficiency," *Quart. J. Econ.*, Feb. 1976, *90*, 75–89.

Implicit Investment Profiles and Intertemporal Adjustments of Relative Wages

By Eric A. Hanushek and John M. Quigley*

The human capital model (see Gary Becker and Jacob Mincer, 1970) is appealing because it introduces a theoretical explanation of earnings differentials that is consistent with rational behavior on the parts of the actors. Nevertheless, the theory is not completely satisfactory because it is built upon unobserved quantities, namely human capital, and the observable implications of this theoretical structure are generally consistent with a variety of other explanations. One objective of this analysis is to consider more directly the implied investment behavior of workers, since it is at that level that human capital theory diverges from alternative theories.

In principle, the relevant tests of the theory relate observations on individual investment activities to earnings patterns over time. However, the absence of direct observation on investment and of longitudinal data on individuals has led to behavioral models developed for and analyzed using cross-sectional data. The specification of the model in cross-sectional terms, however, requires a number of very strong assumptions and precludes estimation of many key parameters of the underlying model. Thus our second objective is to expand the conceptual model to address intertemporal dynamics and to analyze short-run variation in the returns to human investment.

Finally, the stability of earnings profiles over time is a subject of interest in itself. A number of past studies of the rates of return to schooling indicate some instability in the estimates when they are made at different points in time (see Becker, Richard Freeman, Giora Hanoch, and W. Lee Hansen). Such intertemporal differences are often explained by appeal to some sort of aggregate adjustment over a particular interval (see Freeman and Anders Klevmarken). However, the exact nature of these changes and their relationship to rates of return on human capital investment for different subpopulations are never clearly articulated.

This paper begins by extending the model of human capital accumulation to distinguish between the effects of labor market experience and aging on observed earnings profiles, to incorporate other individual differences explicitly, and to incorporate short-run dynamics. On this basis, the implied postschool investment profiles are then estimated, and the impact of aggregate economic conditions on earnings profiles is considered.

I. The Conceptual Model

Typically, human capital models applied to earnings differences begin with a simple description of the investment behavior of a single individual, and then make a series of strong assumptions about the homogeneity of individuals and the pattern of dynamic economic changes so that empirical tests can be conducted with a cross section of individuals. While several problems with this research strategy have been noted (see Alan Blinder and Mincer, 1974), such highly stylized models have led to empirical analyses in which the underlying investment parameters are generally unidentified. In particular, empirical models often estimate earnings as a function of schooling and some transformation of age—a model which is consistent with many "stories" about the labor market, not just a model of human

*Institution for Social and Policy Studies and department of economics, Yale University. Financial support for this research was provided under grants SOC74-245676 and SOC74-21391 from the National Science Foundation and by a grant from the Sloan Foundation. Research assistance was provided by J. A. Ahlstrom. The paper benefited from the comments of Gary Fields.

capital investment. This section expands the basic conceptual framework and demonstrates how information about the implied investment schedules and about labor market dynamics can be unraveled.

In its simplest static form (for example, see Mincer, 1974), the human capital model postulates that

$$(1) \qquad E_n = E_{n-1} + r_{n-1}C_{n-1}$$

where E_n is the potential earnings of an individual with n years of experience (i.e., the earnings obtainable if no resources are used for investment in human capital in the nth year); C_{n-1} is the amount of human capital investment in the $n - 1$ year of experience, and r_{n-1} is its corresponding rate of return. In terms of an "investment ratio," k_n, at any year of experience (defined as the ratio of gross investment to potential earnings, i.e., C_n/E_n), equation (1) may be represented as

$$(2) \qquad E_n = E_{n-1}(1 + r_{n-1}k_{n-1}) = E_0 \prod_{j=0}^{n-1} (1 + r_j k_j)$$

If we assume that the rate of return is constant ($r_j = r$) and that the terms rk_j are small, a Taylor series approximation of (2) is

$$(3) \qquad log\ E_n \sim log\ E_0 + r\sum_{j=0}^{n-1} k_j$$

If we further assume that all investment comes in the form of reduced earnings, such that $E_n = Y_n + C_n$ where Y_n is observable income, then, by a similar approximation,

$$(4) \qquad log\ Y_n \simeq log\ E_0 + r\sum_{j=0}^{n-1} k_j - k_n$$

Equation (4) states that the observed earnings of a worker with n years of experience are related in a simple way to his initial potential earnings, the return on all previous capital investments, and his investment decision in year n.

The exact specification of the investment profile is crucial to any empirical analysis. For a variety of well-known reasons (see Yoram Ben-Porath, 1967, 1970; Blinder and Yoram Weiss; William Haley; James Heckman; Sherwin Rosen), the rational investor will reduce his human capital investment over his lifetime, but the exact shape of the profile cannot be deduced from theory.[1] A common thread in the conceptual discussions is that postschooling investment should be related to the labor market experience of the individual. In this analysis, we assume that the investment ratios (k_n) for postschooling investments decline linearly with actual labor force experience. While this precise form is not derived from a particular lifetime utility-maximization model, it has been shown in at least one recent study (see Klevmarken and Quigley) to be a reasonable empirical approximation to a completely general lifetime investment schedule; it offers the considerable advantages of parsimony and tractability in empirical work; and it has been used quite commonly in past research (see, for example, Mincer, 1970).

To see the implications of this assumed investment path, let n = years since completion of schooling (potential experience), and λ^i_m = the proportion of the mth elapsed year spent in work activity by individual i. Then the accumulated experience at the nth year of potential experience, a^i_n, is

$$(5) \qquad a^i_n = \sum_{m=0}^{n-1} \lambda^i_m$$

and the assumed investment path is given by

$$(6) \qquad k_j = A + Ba^i_j$$

with $0 < A < 1$ and $B < 0$

Even with this simple investment schedule, calculation of total investment (for use in (4)) generally requires knowledge of the entire past history of labor market experience. However, this information is not required in two specific cases. If all individuals are fully employed in all years after school ($\lambda^i = 1$), actual experience equals potential

[1]Theoretical results depend crucially upon the production function for human capital, for which there is no evidence. See, for example, Haley.

experience (a linear transformation of age), and substitution into (4) indicates that the *log* of observed earnings is a quadratic function of potential experience. This has become a quite common specification of earnings functions, largely because many convenient data sets provide information about workers' ages, but not about their labor market experiences. Unfortunately the effects of aging and labor market experience are completely intertwined, and the parameters of interest (A, B, and r) are unidentified.[2] The only refutable hypotheses in this form are that the coefficients on the linear and squared potential experience terms are positive and negative, respectively. But few people would believe that the signs of these coefficients offer a very powerful test of the validity of the underlying model.

While complete data about the profile of individual labor market experiences may be unavailable, information about total labor market experiences can be incorporated into this model in a more general case. If each worker devotes a fixed proportion of each year to labor market activities ($\lambda'_j = \lambda' = a'_n/n'$), the resulting specification of (4) in terms of observable variables becomes

$$(7)\quad log\ Y'_n = log\ Y_0 + [rA]n' + \left[\frac{rB}{2}\right]n'a'_n - \left[\frac{rB}{2} + B\right]a'_n$$

Importantly, by incorporating information about cumulative labor market activities, the underlying parameters can now be identified, even in a single cross section, and the implied investment profiles can be estimated.

A second concern with this model is the static assumption underlying conventional analysis. There are exogenous changes in available capital, productivity, and organization; more generally, the fruits of economic growth are certainly reflected in the time path of earnings received by workers, as are short-run shifts in the supplies and demands for workers in given educational levels. This implies that the lifetime income profile of a cohort of identical workers will differ from (and in general be less concave than) the cross-sectional earnings pattern of otherwise identical workers with differing experience levels.

To recognize these dynamic factors, assume that the earnings profiles of a particular group (say, individuals in the same race/sex/schooling class) are shifted proportionately as a result of changes in productivity, organization, etc., and in response to short-run excess demands for a particular class of labor. Let $\exp\{b_m\}$ be the growth of earnings due to these factors at year m; this yields

$$(8)\quad \begin{aligned} E_{n,t} &= E_{n,t-1}\exp\{b_{t-1}\} \\ &= E_{n,p}\exp\left\{\sum_{m=p}^{t-1} b_m\right\} \\ &= E_{n,p}\exp\{\beta_{t-1,p}\} \end{aligned}$$

where $\sum_{m=p}^{t-1} b_m = \beta_{t-1,p}$

and the second subscript on earnings signifies real time.

Together, (7) and (8) provide a model of the systematic variation in the earnings of a representative individual over his lifetime. However, the potential earnings of individuals with identical schooling, age, and experience clearly vary due to other systematic as well as random influences. We can describe differences in individual profiles as

[2]The resultant model in terms of observable variables is:

$$log\ Y_n = log\ Y_0 + r\left(A - \frac{B}{2} - \frac{B}{r}\right)n + \frac{rB}{2}n^2$$

One way of identifying the underlying parameters is the following: assume that $k_t = 1$ during schooling and the investment function in equation (6) begins at the end of schooling; then it is possible to separate out a term rS where S is years of schooling. If the rate of return on schooling equals the rate of return on postschool investments, the other parameters of the investment function can be identified (see Mincer, 1974). However, there is little a priori reason to assume that these returns are the same. Further, since it appears that the return to schooling varies by schooling level (see Hanoch), the assumption is open to more serious question— since it is not even clear which schooling rate of return should be used.

$$(9) \qquad E^i_{0,t} = E_{0,t} \exp \{X^i\gamma + \mu^i\}$$

where $E^i_{0,t}$ is the potential earnings of an individual in a given schooling class at entry into the labor force, X^i is a vector of systematic differences in earnings profiles due to other measured characteristics of individuals (such as measures of ability or physical handicap), and μ^i is an individual-specific stochastic term which represents a composite of unmeasured attributes (such as quality differences in schooling, attitudes, motivations, or pure luck) which affect the individual's earnings over his lifetime. In addition to these predetermined factors, an individual's observed earnings at any point may depart from his own profile as

$$(10) \qquad Y^i_{n,t} = \tilde{Y}^i_{n,t} \exp \{\nu^i_t\}$$

where $\tilde{Y}^i_{n,t}$ is the expected earnings for an individual and ν^i_t is a stochastic term.

Combining these aspects of earnings yields a testable form of the human investment model which depicts the dynamic earnings path of a class of individuals:[3]

$$(11) \quad \log Y^i_{n,t} = \log \tilde{Y}^i_{0,p} + \beta_{t-1,p} + [rA]n^i + \left[\frac{rB}{2}\right] n^i a^i_n - \left[\frac{rB+B}{2}\right] a^i_n + X^i\gamma + \mu^i + \nu^i_t$$

As specified in equation (11), the logarithm of earnings for an individual of a given schooling class is linearly related to his potential experience (n, the number of elapsed years since school completion), his total labor market experience (a_n, the accumulation of the time spent actually working), and their interaction ($a_n n$), corrected for the vector (X) of systematic differences in ability, etc., and for intertemporal shifts (β) in the earnings profile of his class. The model includes an individual specific error term as well as a more conventional stochastic term which varies by individual and time period. The full model in (11) can be estimated from panel data on individuals that includes measures of earnings or wage rates, schooling, age, and actual labor market experience. If we assume that these errors are normally distributed with zero mean and

$$\begin{aligned} E(\mu^i\mu^j) &= \sigma^2_\mu \quad \text{if} \quad i = j \\ &= 0 \qquad\quad i \neq j \\ E(\nu^i_t\nu^j_\tau) &= \sigma^2_\nu \quad \text{if} \quad i = j, t = \tau \\ &= 0 \quad \text{otherwise} \end{aligned}$$

then the specification follows the general error components form (see T. D. Wallace and A. Hussain).[4]

The model specification in (11) uses differences across individuals in the accumulation of labor market experience to provide information about the underlying investment profiles and the rates of return to postschool investments. Further, it provides additional information about how earnings profiles shift over time. Finally, the stochastic specification permits a direct test of the importance of unmeasured in-

[3]A series of cross sections in which individuals are not linked is sufficient for estimating all but the error components structure; a single cross section implies the β's and error components cannot be estimated; and missing information about actual labor force experience implies that only the model in fn. 2 can be estimated. Conceptually, two considerations define the appropriate stratifications for analysis: the investment schedule considers only postschooling investment and the dynamics relate to shifts for homogenous groups. Therefore, by stratifying into schooling groups, in empirical analysis, earnings differences due to schooling differences are implicitly included in different intercepts. Important dynamic changes include differences among schooling/race/sex groups.

[4]The error components model, which provides more efficient estimates, is estimated by generalized least squares. This specification assumes that the individual component μ^i for different individuals is drawn from a common distribution with a mean of zero. In this specification, the individual component is fixed. This may be unrealistic over a lifetime but is a reasonable approximation over a limited period of time such as the seven years of data used in this work; empirically this specification was found to be virtually identical to estimates which also allow for individual serial correlation in the errors (see Lee Lillard and Robert Willis). Additionally, this specification assumes that the error components (μ^i and ν^i_t) are independent of each other and of the exogenous variables in the model. Finally, the actual estimation procedure differs from that in Wallace and Hussain because all the individuals are not observed for the same number of time periods; this requires a correction in the estimation of ρ, the proportion of residual variance arising from unmeasured differences in individuals.

dividual differences which systematically affect earnings relative to those factors postulated by economic theory.

The appropriate measure for the dependent variable in the theoretical analysis is somewhat unclear. Most previous analyses of human capital have used annual earnings (i.e., total income from labor) as the measure of the dependent variable. To be sure, due to problems of data availability, this has not always been a free choice, yet the implications of this choice are seldom discussed.

There is a fundamental distinction between physical capital and its counterpart imbedded in human beings. The wage rate plays a dual role; it simultaneously represents the return on human capital and the price at which work is substituted for leisure. Because work presumably involves some disutility, the amount of investment in human capital can, indeed will, affect its rate of utilization.

If an individual varies his labor supply as a function of his wage rate, an analysis of annual earnings could give quite misleading impressions of the returns to human capital. The rate of return is calculated with respect to an unobserved stock of human capital, so a given annual earnings can be consistent with either a large stock of capital and a low utilization rate or with a small stock of capital utilized intensively.

Alternatively, individual wage rates can be used as the dependent variable. The theory requires measuring the increase in productivity that is associated with increased investment. The wage rate (in a competitive economy) may be interpreted simply as the productivity associated with a given stock of human capital at a standard utilization rate. This measure of productive capacity seems more within the spirit of the capital investment model, and in the empirical section below we emphasize wage rates. In principle, however, joint estimation of labor supply and wages would be still preferable.[5]

II. Empirical Results

The models described above are estimated using panel data from the Michigan Panel Study on Income Dynamics, which provide annual earnings, wages, and information on personal characteristics for individuals in a sample of about 5,000 households during the period 1968–74. The analysis presented here concentrates upon males, age 16–60 in 1968.[6]

Tables 1 and 2 present estimates of the parameters of equation (11) separately for white and black workers of three schooling classes (0–8 years, 9–12 years, and 13 or more years), with the *log* of wages as the dependent variable.[7] For each stratification, dummy variables are included to reflect the completion of a particular level: at least six years; high school graduation; college graduation; postgraduate education. Three

[5]Similar considerations, but a different standardization, are noted in Richard Eckaus. Using wage rates does neglect any possible effect of utilization on productivity and could understate the private (but *not* social) return on investments to the extent that part of the returns come from less involuntary unemployment. Theoretical developments of models with endogenous wage rates and labor supply require far more stringent assumptions to arrive at an analytical solution (see Heckman and Haley) and are yet to be developed to a point of empirical usefulness. In the empirical section below, we report results using wage rates. We also note any differences when annual earnings are used. Results estimated from annual earnings are available on request.

[6]The analysis is confined to males since the labor force participation decision cannot reasonably be considered exogenous for females (as we implicitly assume). Individuals at or nearing retirement were also excluded. There were 2,793 males in the original sample who met the age restrictions and who were not retired during the sample period. A total of 372 of these were eliminated from the sample because they were neither white nor black or because other data were missing. The sample includes 1,766 white workers and 655 blacks. Individuals are included in the sample only if they had positive earnings for at least three of the possible seven years. The average number of time-series observations for individuals in the sample is 6.27 out of a possible 7 periods.

[7]The average hourly wage rate was estimated by dividing gross labor earnings by the reported (annual) hours worked. There is obviously a large component of measurement error in this computation. However, as compared with models of annual earnings (see fn. 5), the estimated coefficients are generally more precise and the overall explanatory power of the models is slightly higher.

TABLE 1—WAGE MODELS FOR WHITE MALES[b]

Variable	Elementary School	High School	College	Pooled
Schooling Completion				
6-8 years	−.0854			−.0138
	(.8746)			(.1518)
12 years		.1098		.1249
		(3.4772)		(1.3206)
16 years			.2599	.2765
			(6.9561)	(2.8173)
More than 16 years			.3732	.3922
			(7.3240)	(7.7868)
Other characteristics: X^i				
Ability score	.0022	.0350	.0224	.0245
	(.1819)	(4.4961)	(2.1016)	(4.3959)
Health Limitation	−.0321	−.0636	.0026	−.0394
	(1.0274)	(3.3968)	(.0975)	(2.8758)
South	−.1048	−.0372	−.0045	−.0357
	(3.1968)	(2.1936)	(.2018)	(2.8849)
Interaction $a_n^i \cdot n^i/10^2$	−.0768	−.0857	−.1159	−.0923
	(5.1231)	(11.9484)	(11.5974)	(18.7482)
Actual Experience: $a_n^i/10$	.4909	.3253	.4260	.4085
	(4.9270)	(4.7890)	(4.9267)	(8.9660)
Potential Experience: $n^i/10$	−.0961	.1305	.1740	.0889
	(1.1399)	(2.0532)	(2.1455)	(2.1078)
Intercept				
Elementary school	.6205			.1554
	(3.1943)			(1.4629)
High School		.2102		.1184
		(2.5292)		(1.2815)
College			.4532	.3316
			(3.8246)	(3.4570)
β_{69}	.1089	.0660	.0899	.0798
	(3.529)	(4.102)	(4.193)	(6.7154)
β_{70}	.2090	.1450	.1379	.1520
	(6.737)	(9.088)	(6.525)	(6.2636)
β_{71}	.2518	.2052	.2012	.2110
	(7.988)	(12.873)	(9.527)	(12.0500)
β_{72}	.3515	.2884	.2388	.2811
	(10.882)	(17.973)	(11.284)	(11.3504)
β_{73}	.3334	.3620	.3294	.3505
	(9.400)	(21.691)	(15.041)	(16.8503)
β_{74}	.4599	.4295	.3838	.4209
	(12.304)	(24.644)	(16.864)	(19.1455)
R^2	.1716	.2526	.2503	.2496
ρ	.6432	.6394	.5904	[a]
σ^2	.3240	.2430	.2729	[a]
σ_μ^2	.2080	.1550	.1630	[a]
σ_ν^2	.1160	.0880	.1100	[a]
Number of observations	1685	5483	4029	11197
Number of individuals	251	873	642	1776

[a]Equations were pooled using the values of ρ estimated for each of the subpopulations.
[b]*The total error variance is decomposed into an individual specific and a purely random component.* These estimated error components are subsequently used in a generalized least squares estimation procedure (Wallace and Hussain). The components are defined as follows: σ^2 = estimated total error variance; σ_μ^2 = estimated individual specific error variance, i.e., variance of μ_i; $\sigma_\nu^2 = \sigma^2 - \sigma_\mu^2$ = purely random error variance; and ρ = proportion of total error variance which is individual specific = $\sigma_\mu^2/\sigma^2 = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\nu^2)$. t-statistics in parentheses.

TABLE 2—WAGE MODELS FOR BLACK MALES[b]

Variable	Elementary School	High School	College	Pooled
Schooling Completion				
6-8 years	−.0454			−.0621
	(.5985)			(.9026)
12 years		.1795		.2214
		(4.3839)		(2.9231)
16 years			.1644	.1741
			(1.3837)	(1.2887)
More than 16 years			.6443	.5148
			(4.9253)	(4.2544)
Other characteristics: X'				
Ability score	.0369	.0203	.0211	.0279
	(2.9296)	(2.6110)	(1.2224)	(4.5541)
Health Limitation	−.1102	−.0558	.0515	−.0599
	(2.6174)	(1.7110)	(.6406)	(2.4444)
South	−.1341	−.0637	.0268	−.0676
	(2.7047)	(2.5029)	(.3983)	(3.2233)
Interaction $a_n^t \cdot n^t / 10^2$	−.0553	−.0614	−.1158	−.0554
	(3.1456)	(4.7680)	(4.0981)	(6.5405)
Actual Experience: $a_n^t / 10$	.2782	.3373	.3920	.2764
	(3.2805)	(4.4035)	(1.1048)	(5.5906)
Potential Experience: $n^t / 10$	.0418	.0044	−.0001	.0194
	(.6894)	(.0665)	(.0003)	(.4718)
Intercept				
Elementary school	−.0331			.1097
	(.1893)			(1.2020)
High School		.2680		.1044
		(3.3030)		(1.4646)
College			.3942	.2736
			(1.9563)	(3.1079)
β_{69}	.1941	.0973	.0721	.1334
	(5.211)	(3.211)	(0.832)	(5.8636)
β_{70}	.2651	.1614	.1783	.2027
	(7.061)	(5.433)	(2.129)	(4.1815)
β_{71}	.3716	.2399	.3328	.2974
	(9.778)	(8.192)	(4.014)	(6.3771)
β_{72}	.4759	.3404	.3530	.3924
	(12.208)	(11.706)	(4.366)	(8.1591)
β_{73}	.4172	.3887	.5343	.4328
	(7.632)	(12.397)	(6.050)	(8.0079)
β_{74}	.5133	.4514	.6665	.5125
	(8.725)	(13.026)	(6.951)	(9.1087)
R^2	.2227	.2434	.2899	.2487
ρ	.5613	.4708	.3224	a
σ^2	.3257	.2101	.2273	a
σ_μ^2	.1830	.0990	.0730	a
σ_ν^2	.1430	.1110	.1540	a
Number of observations	1448	2148	390	3986
Number of individuals	222	366	67	655

[a,b]See Table 1; *t*-statistics in parentheses.

additional variables are also included in the regressions to reflect: "native ability," as measured by the score on a short sentence test administered in 1972; "health limitations" affecting employment, as self-reported for each year; and residence in the South for each year. Each regression also includes a measure of potential experience, actual experience, and an interaction term.[8]

The regression results pooled across schooling groups indicate that the model explains about 25 percent of the variance in logarithmic wages for both black and white workers. The results in Tables 1 and 2 explain between 17 and 29 percent of the variance in *log* wages within race/schooling classes. The combined explanatory power of the stratified models for whites is 28 percent (20 percent from within schooling group regressions and 8 percent from stratification); for blacks the comparable figure is 26 percent (18 percent within group and 8 percent between groups).[9]

The estimates suggest that there are substantial returns to graduation for those with some high school or college training. The wages of white high school graduates are about 11 percent higher than those of otherwise comparable nongraduates; for blacks the estimate is 18 percent. The coefficients imply that the wages of white (black) college graduates are about 26 percent (16 percent) higher than those of college drop-outs; for those with postgraduate education, wages are higher by an additional 37 percent (64 percent).[10] The variable reflecting elementary school completion is insignificant; however, almost all workers in this schooling group have completed 6–8 years of education.

The other personal characteristics (ability test, health limitations, residence) have the expected signs. The ability measure, although quite crude, has a positive and generally significant effect on wages.[11] There is no discernible positive interaction between schooling and ability, contrary to the findings of John Hause. The difference in earnings associated with a movement from the lowest to the highest ability group ranges from 20 to 75 percent in the different subsamples. Health limitations depress the wages of individuals with a high school education or less by 3 to 11 percent. However, individuals with a college education, generally in less physically demanding occupations, apparently suffer little or no loss with health limitations. Individuals residing in the South tend to have lower wages except for the college educated. The decreasing importance of residence at higher schooling levels is consistent with the notion that labor markets are less regionalized for the more educated (see Hanushek, 1973). The wage difference associated with southern residence is greater for blacks than for whites.

The results indicate quite strongly the importance of actual experience and the interaction term in determining wages, but only for whites is the coefficient of potential

[8]The actual experience of each worker was computed from the answer to the question "How many years of labor force experience do you have?" asked in 1974. Actual experience in prior years was computed by subtracting the cumulative proportion of the years worked from the 1974 figure. For these calculations, full time is defined as 1750 hours/year (or 50 weeks of work at 35 hours/week). Some experimentation was done by defining full time as 1750, 2000, and 2500 hours per year with little effect on the estimates.

[9]An *F*-test rejects the hypothesis of equality of coefficients across schooling groups. These tests allow for intercept and schooling coefficient differences. For whites, $F_{26,11152}$ is 1.97; for blacks, $F_{26,3941}$ is 1.69. ($F(26,\infty) = 1.76$ at the .01 level.)

[10]Note that these are not pure "sheepskin" effects in either the pooled or stratified models. The estimated coefficient represents the returns for completion over the median noncompleter within the same schooling class. For example, the 11 percent return to white high school graduates represents the return to additional years above the median noncompleter in the 9–11 year group plus any sheepskin effect. In the pooled models, separate intercept terms are estimated for the schooling classes, giving the schooling coefficients an identical interpretation, i.e., a 12 percent return to high school graduation as compared to the median noncompleter in this schooling group.

[11]The ability measure is the score on a "short sentence" test. Scores can range between 0 and 13, and it appears that there might be a "topping out" problem in that many people achieve the maximum score. This may lead to the smaller estimated effect of ability for the college groups.

experience consistently significant.[12] (These coefficients are discussed below.)

For white workers, the estimate of ρ (the proportion of residual variation attributable to unmeasured characteristics of individuals, μ^i) suggests that more than half of the residual variation in earnings is individual-specific. For black workers, however, the estimate of ρ is smaller—significantly so for the college group. This arises from a larger transitory component and, perhaps, a smaller permanent component of the error variance for blacks.[13] This finding implies less stability in lifetime wages and earnings for blacks of a given schooling group relative to comparable whites.

It is worth emphasizing the overall results. For both black and white males within any schooling class, the theory of postschool investment explains roughly 25 percent of the variation in productivity; 45 percent is "explained" by other systematic (but unmeasured qualities) of individual workers (for example, their "motivation") or their work histories (for example, their "good luck"); and about 30 percent is completely unexplained.

III. Human Capital Interpretation

The estimates presented in Tables 1 and 2 are based upon a simple model of human capital investment, relating observed hourly wage rates to a linear profile of capital investment and their annual returns. The principal novelty in the analysis is the distinction between the actual labor market experience gained by individuals and the potential experience gained simply by aging. This permits direct estimation of the parameters of the postschool investment profile and the rate of return.[14] Table 3 displays the estimates of these parameters. Part (a) of the table arrays the rates of return to postschool capital investment for each of the six race/schooling groups for male workers. These range from 3.18 to 6.10 percent for wages (and 4.42 to 8.88 percent for earnings (not shown)). The rate of return measured in hourly wages is lower than that measured in terms of earnings, although the pattern of returns across races and schooling groups is generally consistent between wages and earnings. It also appears generally true that postschool investments

[12]Compared with the "conventional" specification, which includes potential experience and its square, the results reported explain between 2 and 9 percent more of the variance in *log* wages and earnings.

[13]A lower estimate of ρ in any stratification will arise from either a small σ_μ^2 or a large σ_ν^2. Because 1,872 families (out of the 4,802 households in our panel begun in 1968) were selected from the Survey of Economic Opportunity (*SEO*) on the basis of low 1966 family incomes, the estimates of σ_μ^2 and ρ could be biased by sample selection. (Of course, if low incomes in 1966 resulted from low transitory components, there would be no bias.) Since the sample of black workers includes a higher proportion of *SEO* families, σ_μ^2 may be biased downwards relative to whites. However, σ_ν^2 is so large that even if the true σ_μ^2 for black workers were as large as the estimates for whites, ρ would still be 8 to 13 percent less for blacks than for whites. It should be noted that the estimation procedure is designed to hold constant any nonrandom factors causing otherwise identical workers to have different permanent incomes; hence any selectivity bias should not affect the estimated coefficients reported in the tables.

[14]Parts of our interpretation of these results would be altered if vintage of schooling also affects earnings (see Weiss and Lillard). Since vintage (school completion date) equals real time t minus potential experience n, the coefficient interpretation depends upon the particular specification of the vintage effects on wages or earnings. For example, if vintage produces parallel shifts in *log* wage profiles, then the rate of return r and the slope of the investment profile B in equation (11) are still identified, even though vintage effects are included in the estimated coefficients of n' and the β's. This identification problem is clearly more important when aging and labor market experience are not separately measured, and has been resolved only by making extremely strong behavioral assumptions (see Weiss and Lillard and Finis Welch). Even when age and experience are separately measured, however, some identification problems remain unless additional information about the specification and measurement of "vintage" is introduced. In principle, several alternatives are possible. Vintage could be parameterized (say by measures of school quality) and incorporated directly into the estimation; or, if vintage could be assumed to cause parallel shifts, then the vintage parameter(s) would appear as an explicit component of the β's (for example, a linear trend in the special case of proportionate shifts) which could be estimated by auxilliary regression on the β's (given sufficient longitudinal information). Unfortunately, currently available data will not support these alternatives.

TABLE 3—ESTIMATED RATES OF RETURN TO POSTSCHOOL INVESTMENT AND IMPLIED INVESTMENT PROFILES

	Elementary School	High School	College	Pooled
(a) Rate of return: r				
Whites[a]	3.18	5.41	5.60	4.63
Blacks[a]	4.05	3.71	6.10	4.09
(b) Slope of investment function: B				
Whites	-.0483	-.0317	-.0414	-.0399
Blacks	-.0273	-.0331	-.0380	. .0271
(c) Intercept of investment function: A				
Whites	(-)	.1825	.3107	.1920
Blacks	1.0321	.1186	(-)	.4743

[a]Shown in percent.

yield higher returns for those with more formal education. In addition, there is some evidence that the marginal return to postschool investment is greater for black workers than for white workers.

It is important to note, however, that the rate of return calculated here is the amount that equates individual earnings along a given profile. Thus, it is a "growth adjusted" rate of return, and it is not the *ex post* rate of return that any individual actually received during the 1968–74 period. The return to postschooling investment (in nominal terms) would be found adding r and the β_t for each year.[15] In real terms, the part of the β_t's that represents price changes (see below) would be eliminated from this calculation.

The slopes of the investment profiles suggest that postschool investment declines at 53 to 140 hours per year, and there is some evidence that investment declines more rapidly for the more educated.[16] However, there is little apparent consistency in the estimates of the intercepts of the investment functions. The instability of these estimates may reflect the nonlinearities of the investment profile at entry into the labor force discerned by other studies (see Klevmarken and Quigley) or, alternatively, vintage effects that are not identified in the models (see fn. 14). Taken literally, the investment profiles imply rather short periods of net investment (for example, less than 6 years for white high school graduates).

IV. The Determinants of Shifts in Profiles

Clearly, the actual returns received by individuals depend upon factors other than their human capital investment strategies, their abilities (even appropriately measured) and their health status. In the long run, they depend upon the amount, type, and quality of complementary capital, its productivity and organization, and other factors. In the short run, the returns received by workers depend upon the excess supplies of workers with particular skills and occupational characteristics.

During the recent period of inflation, recession, and recovery, there has been increasing concern about changes in relative earnings and short-run effects of macroeconomic conditions upon identifiable groups of workers (see Freeman). The β's presented in Tables 1 and 2 indicate that annual changes in profiles for each of the six groups of workers during the 1968–74

[15]If the β's were the same across all groups, then ignoring dynamic factors in a cross-section estimation of the rate of return would not be a serious problem—the real growth rate could simply be added to the estimated rate of return. However, as shown below, it is not possible to assume equal growth rates across groups.

[16]Investment is measured in terms of time equivalents; thus, this is calculated by multiplying the slope of the investment profile times 1750 hours, the assumed full-time work year.

TABLE 4 – REGRESSION COEFFICIENTS FOR THE RELATIONSHIP BETWEEN PERCENTAGE CHANGE IN THE WAGE PROFILE (β_i) AND EXOGENOUS FACTORS

Independent Variable	Linear	Logarithmic
CPI	.642[a]	.874
(1967 = 100)	(5.62)	(6.09)
Real *GNP*	.667[b]	674
(in $B)	(3.03)	(2.72)
Race	.066	.065
(1 = black)	(4.40)	(4.60)
High School	–.063	–.063
(1 = high school)	(–3.82)	(–4.03)
College	–0.61	–.061
(1 = college)	(–3.21)	(–3.43)
Constant	–1.272	–8.671
	(–8.64)	(–7.30)
R^2	.91	.91

Note: t-statistics in parentheses.
[a]times 10^2
[b]times 10^3

period vary significantly. In Table 4, we relate the pattern of these parallel shifts to aggregate economic conditions—as measured by the movements in real output and price levels. Also included are dummy variables for race and schooling groups.

Over 90 percent of the annual shifts in earnings profiles for these groups can be explained by the systematic influence of aggregate conditions (in either linear or logarithmic form) and the three dummy variables.[17] Price changes are not completely passed through into wages (or earnings, not shown). While the price coefficient in the logarithmic models is not significantly different from one, this does indicate potential biases in estimation where wages and earnings are simply deflated by the cost of living. The *GNP* coefficient indicates that a 1 percent increase in *GNP* is reflected in a .67 percent increase in wage profiles.

The results further suggest that—holding investment profiles and other individual characteristics constant—the high school and college groups both lost in relative terms over this period. The marginal returns of the college educated (relative to those who stopped at high school) remained virtually constant.[18] However, the limited time-series information makes it difficult to distinguish between secular changes in relative wages and more short-run phenomena.

Similarly, the position of blacks relative to whites appears to have improved during this period, other things equal. There is an important qualification, however: the racial differences in profiles indicate that the wages of black high school and college graduates would otherwise have fallen over the period, relative to white workers with the same pattern of labor market experience.

V. Conclusions

This paper extends the human capital model by considering the distinction between the actual labor market experience of individuals and their potential experience gained simply by aging. This distinction permits the underlying parameters of the capital investment function and the rates of return to schooling and postschool investment to be estimated directly.

The estimates of the investment model explaining both wages and earnings appear reasonably consistent with the human capital formulation. The implicit rates of return to postschool investment and the slopes of the underlying investment schedules seem plausible. "Growth adjusted" rates of return (those normalized for economic growth, price changes and short-run shifts in labor force demands) range between 3 and 9 percent. The investment profiles themselves (assumed to be linear) decline in a reasonable manner with a rate

[17]The models were estimated by generalized least squares using the 36 estimated β's from Tables 1 and 2. The *GLS* procedure (see Hanushek, 1974) allows for the fact that the β's are themselves regression estimates and thus contain some sampling variation.

[18]These estimates are not, however, directly comparable to Freeman's results, since he explicitly considers "twists" in the profiles (i.e., that the position of young college workers has worsened even though the position of all college workers may not have).

of decline that increases with level of schooling.

There is, however, some reason for caution in interpreting the human capital formulation of the wage and earnings models. First, the intercepts of the investment profile are not well estimated and, taken literally, imply implausibly short periods of positive net investment after leaving school. Second, estimation of different formulations of the same basic model (to explain wage growth as opposed to wage levels) did not yield plausible estimates of the underlying investment parameters.[19] While each of these problems could be explained by nonlinearities in the investment profile, they could also indicate more fundamental problems with the underlying investment model.

The intertemporal shifts in the earnings profiles are explained by *GNP* growth, price changes, and terms relating to specific race and schooling classes. Price changes were not completely passed through to wage changes, indicating potential problems from simply analyzing deflated wages. The elasticity of the profiles with respect to total output is about .67. Over the period 1968 through 1974, the relative wages of blacks has improved while the wages of high school and college educated individuals have fallen relative to those with less education. Differences between high school and college educated workers are insignificant.

[19]Investigation of an alternative specification of (11) produced results which were less consistent with the human capital model. The conceptual framework should, in addition to explaining the level of wages, also explain changes in wages over time. This latter formulation has the advantage that no assumption about the pattern of labor market experiences over a lifetime is required, since the growth in wages will be a function of total experience and the change in experience in two adjacent years. Models of this form did not yield plausible estimates of the investment profile or of the rate of return to postschool investment, perhaps because of problems from estimating wage rates, from the familiar increase in signal to noise in time-series models estimated on differences, or from the increased importance of the assumed linear investment profile. Alternatively, it could reflect more fundamental problems with the human capital model. It is not possible to distinguish among these possible causes.

REFERENCES

Gary Becker, *Human Capital*, New York 1964.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

———, "The Production of Human Capital Over Time," in W. Lee Hansen, ed., *Education, Income, and Human Capital*, New York 1970, 129–54.

A. Blinder, "On Dogmatism in Human Capital Theory," *J. Hum. Resources*, Winter 1976, *11*, 8–22.

——— **and Y. Weiss**, "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ*, June 1976, *84*, 449–72.

R. Eckaus, "Returns to Education with Standardized Incomes," *Quart. J. Econ.*, Feb. 1973, *87*, 121–31.

R. Freeman, "Overinvestment in College Training?," *J. Hum. Resources*, Summer 1975, *10*, 287–311.

W. J. Haley, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–44.

G. Hanoch, "An Economic Analysis of Earnings and Schooling," *J. Hum. Resources*, Summer 1967, *2*, 310–29.

W. L. Hansen, "Total and Private Rates of Return to Investment in Schooling," *J. Polit. Econ.*, Apr. 1963, *71*, 128–40.

E. A. Hanushek, "Efficient Estimators for Regressing Regression Coefficients," *Amer. Statistician*, May 1974, *28*, 66–67.

———, "Regional Differences in the Structure of Earnings," *Rev. Econ. Statist.*, May 1973, *55*, 204–13.

J. Hause, "Earnings Profiles: Ability and Schooling," *J. Polit. Econ.*, May/June 1972, Part II, *80*, S108–38.

J. Heckman, "A Life Cycle Model of Earnings, Learning, and Consumption," *J. Polit Econ.*, Aug. 1976, Part II, *84*, S11–44.

T. Johnson and F. J. Hebein, "Investments in Human Capital and Growth in Personal Incomes 1956–1966," *Amer. Econ. Rev.*, Sept. 1974, *64*, 604–16.

Anders Klevmarken, *Statistical Methods For the Analysis of Earnings Data*, Stockholm 1972.

——— **and J. M. Quigley,** "Age, Experience, Earnings and Investments in Human Capital," *J. Polit. Econ.*, Feb. 1976, *84*, 47–72.

E. Lazear, "Age, Experience, and Wage Growth," *Amer. Econ. Rev.*, Sept. 1976, *66*, 848–58.

L. E. Lillard and R. Willis, "Dynamic Aspects of Earnings Mobility," Nat. Bur. Econ. Res. work. paper no. 150, New York, Sept. 1976.

Jacob Mincer, "The Distribution of Labor Incomes: A Survey," *J. Econ. Lit.*, Mar. 1970, *8*, 1–26.

———, *Schooling, Experience, and Earnings*, New York 1974.

S. Rosen, "Human Capital: A Survey of Empirical Research," work. paper no. 76-2, Univ. Rochester, Jan. 1976.

———, "A Theory of Life Earnings," *J. Polit. Econ.*, Aug. 1976, Part II, *84*, S545–67.

T. D. Wallace and A. Hussain, "The Use of Error Component Models in Combining Cross Section with Time Series Data," *Econometrica*, Jan. 1969, *37*, 55–72.

Y. Weiss and L. Lillard, "Experience, Vintage, and Time Effects in the Growth of Earnings: American Scientists, 1960–1970," Nat. Bur. Econ. Res. work. paper no. 138, New York, May 1976.

F. Welch, "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *63*, 893–907.

Study Research Center, *A Panel Study of Income Dynamics: Study Design, Procedures Available Data*, Ann Arbor 1972.

U.S. Office of Economic Opportunity, "Survey of Economic Opportunity," (*SEO*) conducted spring 1967, available on tape, Data Bank, Univ. Wisconsin.

Product Safety: Liability Rules, Market Structure, and Imperfect Information

By DENNIS EPPLE AND ARTUR RAVIV*

The problem of product safety has been a subject of growing concern in recent years. Recognition of the enormity of loss to both person and property arising from product failure has brought a reconsideration of public policies. In particular there has been a trend toward greater regulation of hazardous products. In addition, judicial decisions have increasingly placed liability for failure on the producer of the faulty product. From an economic standpoint the interesting questions are: what are the effects of various institutional arrangements on the safety of products; and which arrangements result in the production of socially optimal goods under alternative assumptions about market structures, availability of insurance, and consumers' information about product safety characteristics?

The first major attempt to elucidate the problem was provided in a series of papers by Roland McKean, James Buchanan, Guido Calabresi, Robert Dorfman, and others.[1] These papers identify the legal issues and begin to develop a framework for economic analysis of the problem. Two attempts at a more formal analysis were provided by John Brown and Walter Oi. Brown considers the assignment of liability for failures of an asset with random life. He focuses exclusively on the demand side taking product characteristics as exogenous. Oi analyzes both the consumption and production of an unsafe product under conditions of perfect competition, risk neutrality, and perfect information. His primary focus is on the desirability of product safety regulation. He concludes that governmental intervention will generally result in a reduction of consumer welfare. Oi's analysis is criticized by Victor Goldberg who argues that the major consideration in determining policy regarding product safety is imperfect consumer information. However, he does not develop an alternative formulation of the problem. Analysis of another aspect of the problem has been provided by Koichi Hamada who investigates the effect of liability rules on income distribution.

The previous analyses employ several simplifying assumptions which limit their applicability in resolving questions of policy regarding product safety. In the analysis that follows we develop a more general model which incorporates several important features. First, product safety characteristics are variable and are determined from an equilibrium analysis. Second, we clearly distinguish between two product safety characteristics, the probability of failure and the severity of the damages. Third, we analyze durable goods considering the multiperiod aspects of the problem arising from the possibility of a sequence of failures followed by replacements. Fourth, we analyze product safety under different market structures. Fifth, we consider the effects of imperfections in external insurance markets on product safety. Sixth, we explicitly allow for imperfect information by postulating a probability distribution reflecting the consumer's subjective judgment about product characteristics.

In this paper we consider a durable good which fails randomly. Failure of the good results in a loss which consists of damages and possibly the destruction of the good itself. The safety characteristics are determined by the manufacturer, who takes account of the consumer's behavior. We determine the effects of market structure and liability rules on the chosen characteristics

*Assistant professors, Graduate School of Industrial Administration, Carnegie-Mellon University. We wish to acknowledge helpful comments from John P. Brown, Tim McGuire, Tom Romer, Michael Visscher, and Allan Zelenitz.

[1] These papers are published in a symposium entitled: "Products Liability: Economic Analysis and the Law" in the *Univ. Chicago Law Review*, Fall 1970.

of the good. Clearly the safety characteristics of the good affect its cost of production and therefore the price paid by consumers. Thus our analysis of the desirability of alternative liability rules is based on the determination of their effects on consumer welfare. We show that product safety and consumer welfare depend on the terms of available insurance contracts. We also show that consumer's information plays a major role in determining the safety characteristics of the product and thereby consumer welfare. The desirable liability rule is shown to depend on the amount of information available to consumers.

While we do not claim that our model embodies all features of the problem of product safety, it does shed light on several controversial issues which thus far have been debated without benefit of a formal model. For example, the National Commission on Product Safety proposed regulating or banning "unreasonably" hazardous products without showing that such policies will either improve product safety or increase consumer welfare. On the other hand, economists (see Buchanan and Oi) have tended to endorse caveat emptor as the preferred liability assignment and have argued that governmental intervention will eliminate certain classes of products which are socially desirable.

The demand for the risky product and insurance against potential losses are considered in Section I. The analysis is in a multiperiod framework and extends previous work by Brown. Section II includes the equilibrium determination of product safety characteristics under alternative market structures and liability rules in the presence of full insurance. The assumption of full insurance is relaxed in Section III. Section IV is devoted to the analysis of alternative liability rules when consumer information is imperfect. Conclusions and discussion are contained in Section V.

I. Consumer Demand

In this section we analyze the demand for a risky product. We assume that the product is purchased for a price P and may fail in any one of T periods. The probability of failure θ is constant each period and independent of the outcome in the previous periods. When the product fails it causes a damage of size L and in addition must be replaced for the purchase price P. The properties of the distribution of product failure are the same as the properties obtained if the time until failure is exponentially distributed. The exponential distribution is commonly used in the literature on product reliability, and there is considerable evidence (see Richard Barlow and Frank Proschan, p. 18) that it is a good approximation for a wide range of products.

We postulate the existence of an insurance market where the consumer can insure against both the loss caused by the product and the loss of the product itself. Following the insurance literature (see Kenneth Arrow) we assume that the premium charged by the insurance company is proportional to the expected value of claims. The premium is paid at the outset for coverage through time T. Denoting the proportionality factor by λ, the discount factor by β, the insurance premium by R, and the amount that the insured receives in the event of failure by I, we obtain:

$$R = \lambda \sum_{t=1}^{T} \beta^t \theta I \tag{1}$$

We assume that $\lambda \geq 1$. When $\lambda = 1$ the premium equals the expected value of claims and insurance is actuarially fair. The case of fair insurance will be of special importance in what follows. For simplicity, we assume that the amount of insurance purchased I is constant through time.

Since our discussion is directed toward major assets, we assume that each consumer purchases at most one unit of the asset. The consumer who does purchase the asset is assumed to maximize his expected utility by choice of insurance and the amount of other goods consumed each period over a planning horizon of T periods. That is, the consumer seeks to maximize:

$$(2) \qquad E\sum_{t=0}^{T} \gamma^t U(C_t)$$

where C_t is the amount spent on other consumption goods in period t. Implicit in the above expression is the assumption that the utility function is separable through time and is discounted at the rate γ. We assume that U is increasing, strictly concave in its argument, and that marginal utility approaches infinity as consumption approaches zero. The above optimization is subject to the constraint that the present value of total expenditures cannot exceed the initial wealth W_0. Our assumptions regarding the utility function assure that consumption is positive in each period. We assume that the consumer can both borrow and lend at the market discount factor β.

The sequence of decisions is as follows: at the beginning of period zero the consumer decides on the amount of insurance to purchase and on the amount of consumption in period zero. The decision on the amount to consume in each subsequent period is made at the beginning of the period contingent on the wealth remaining at that time. At the end of each period through T − 1, the consumer learns whether the asset has failed. If it has, he pays the uninsured loss $P + L - I$ at that time. In period T all remaining wealth is consumed.

Wealth at the beginning of each period W_t, can now be related to wealth at the beginning of the preceding period W_{t-1}, given the consumption decision in the preceding period and the information concerning failure of the asset. Since the replacement cost and the damage appear symmetrically, we will denote the total loss by $A = P + L$.

$$(3a) \qquad W_1 = \begin{cases} \frac{1}{\beta}(W_0 - C_0 - R) - (A - I) & \text{with probability } \theta \\ \frac{1}{\beta}(W_0 - C_0 - R) & \text{with probability } 1 - \theta \end{cases}$$

For t = 2, 3, ..., T

$$(3b) \qquad W_t = \begin{cases} \frac{1}{\beta}(W_{t-1} - C_{t-1}) - (A - I) & \text{with probability } \theta \\ \frac{1}{\beta}(W_{t-1} - C_{t-1}) & \text{with probability } 1 - \theta \end{cases}$$

Notice that in (3a) we implicitly assumed that the initial wealth W_0 is net of the initial purchase price of the asset.

The consumer maximizes (2) subject to the wealth equations (3). The solution is via stochastic dynamic programming and is presented in the Appendix. It is proved there that full insurance is optimal if and only if insurance is actuarially fair. We also show that if the discount rate for utility γ equals the market discount factor β, and insurance is actuarially fair, then consumption is equal in all periods.

The above result contradicts the conclusion previously derived by Brown in an analysis of a similar problem. Brown concluded that less than full insurance was optimal even when insurance is actuarially fair. There are two shortcomings to his formulation. First, the consumption path is not determined in the model. Second, both losses and the premiums are borne instantaneously rather than being spread over time, that is, all payments are instantaneously subtracted from an exogenously determined "normal level of consumption" (Brown, p. 150). In our model both the consumption decision and the insurance purchasing decision are endogenous. As we have shown in the Appendix, when the consumer is able to adjust his consumption he will spread his premium charges through time, and he will fully insure. It was also shown that if $\beta = \gamma$ (as is implicit in Brown's derivation) consumption after payment of the insurance premium is equal in all periods in contrast to Brown's assumption.

When insurance is actuarially fair, all uncertainty inherent in owning a risky asset is eliminated by insurance. On the other hand, if insurance is loaded, the optimal level of coverage depends on the utility function,

the initial wealth of the individual consumer, and the insurance loading. As a result, it is extremely difficult to characterize the total cost of owning the good since the consumer chooses to bear part of the risk of owning the good. Therefore, in our analysis in the following section we impose the assumption of full insurance even when $\lambda > 1$. The case of optimally chosen partial insurance will be taken up in Sections III and IV.

To analyze product safety we must characterize the aggregate demand function for the product. *Under full insurance*, the total cost of owning the good for T periods is made up of the purchase price P and the insurance cost R. We denote the total cost by

$$(4) \quad H = P + R = P + \lambda\theta A \sum_{t=1}^{T} \beta^t = P + \lambda\theta AB$$

where we have used the definition of R evaluated at $I = A$ and B denotes the present value of an annuity of $1 per period for T periods. Recall that at the outset we have assumed that each individual purchases at most one unit of the asset. The purchase decision is based on the full cost of owning the asset for T periods which is given by (4). Since different individuals have different reservation prices for this asset, the aggregate demand function is downward sloping. Thus we assume

$$(5) \quad Q = Q(H) \qquad Q'(H) < 0$$

II. Market Equilibrium Under Full Insurance

In this section we consider product safety under alternative liability rules and market structures. Product safety will be reflected both in the frequency of failure and in the size of the loss if failure occurs. While both dimensions of product safety will generally be imbedded in a given product, we find it useful to maintain the conceptual distinction. An example will clarify this distinction. Consider the steering mechanism of an automobile. The probability of failure of the mechanism will depend on the design and quality control exercised in production. Clearly this probability can be reduced by an increased effort on the part of the producer. On the other hand, many automobile accidents occur due to events beyond the control of the manufacturer. Although he cannot control the probability of the accident, his decisions do affect the magnitude of the loss through the design of bumpers, air bags, etc. We shall investigate the determination of these two characteristics of the good in parts A and B of this section. In each of these subsections we analyze two alternative market structures and two liability rules for each market structure. The results are summarized in C.

On the supply side we assume the unit cost of production is

$$(6) \quad C(\ ), \qquad C'(\ ) < 0, \qquad C''(\ ) > 0$$

The argument of the cost function will be L in part A and θ in part B. As indicated in (6), we assume constant returns to scale with unit cost a decreasing, strictly convex function of the severity of the loss or of the frequency of failure.

Under *consumer liability* the individual purchases a hazardous asset for price P and fully insures against potential losses. As described in Section I, the aggregate demand for the product depends on H as shown in (5). The producer maximizes expected profit which is the sum of profit from initial sale of the asset and profit from replacement when failures occur.

$$(7) \quad \pi = [P - C(\ )]Q(H) + \theta\sum_{t=1}^{T} \beta^t[P - C(\ )]Q(H) = [P - C(\ )]\, Q(P + \lambda\theta(P + L)B)(1 + \theta B)$$

When the *producer is liable* the consumer purchases the product for price H and the producer bears all costs associated with failure of the good. The objective function of the producer is

$$(8)\quad \pi = [H - C(\ )]Q(H) - \lambda\theta[C(\ ) + L]Q(H)B = Q(H)[H - C(\ )(1 + \lambda\theta B) - \lambda\theta LB]$$

The first term in (8) is the profit from the initial sale and the second term is the producer's cost of insuring against losses arising from failures of the product. We assume that the loading represents the real cost of insurance, and this cost is incurred whether the producer self-insures or purchases the insurance from a third party.

Note that under consumer liability, H is the sum of the purchase price and the insurance premium paid by the consumer to a third party insurer. Under producer liability it is the full price paid for a product which carries a complete guarantee.

A. *Determination of the Severity of Loss*

Severity of the loss and consumer welfare will now be determined for four cases corresponding to the alternative liability rules and market structures:

CASE 1: *Consumer Liability—Monopoly*

From (7), the first-order conditions determining P and L for a monopolist are

$$(9)\quad Q + (P - C)Q'(1 + \lambda\theta B) = 0$$

$$(10)\quad -C'Q + (P - C)Q'\lambda\theta B = 0$$

By substituting (9) into (10) the condition for L is

$$(11)\quad C'(1 + \lambda\theta B) + \lambda\theta B = 0$$

CASE 2: *Consumer Liability—Perfect Competition*

The competitive firm treats H as given because the consumer is indifferent between different values of P and L which leave H constant. Solving (4) for P and substituting into (7) we obtain

$$(12)\quad \pi = Q(H)[H - \lambda\theta LB - (1 + \lambda\theta B) \cdot C(L)](1 + \theta B)/(1 + \lambda\theta B)$$

The two conditions to determine the optimal values P and L for the competitive firm are the zero profit condition and the requirement that L maximize (12) when H is treated as a parameter:

$$(13)\quad P - C = 0$$

$$(14)\quad C'(1 + \lambda\theta B) + \lambda\theta B = 0$$

Clearly condition (14) is equivalent to choosing L to minimize total cost of owning the asset: $C(1 + \lambda\theta B) + \lambda\theta LB$. By comparing (11) to (14), it is apparent that the chosen L is independent of market structure.

CASE 3: *Producer Liability—Monopoly*

The monopolist's first-order conditions are obtained by differentiating (8) with respect to H and L:

$$(15a)\quad Q + Q'[H - C(1 + \lambda\theta B) - \lambda\theta LB] = 0$$

$$(15b)\quad C'(1 + \lambda\theta B) + \lambda\theta B = 0$$

CASE 4: *Producer Liability—Perfect Competition*

For a competitive firm H is taken as given and the equilibrium conditions determining H and L are zero profit and a condition identical to (15b). Therefore it follows immediately that under producer liability, L is the same for either market structure.

We have shown above that L is independent of market structure under each liability rule. Moreover, condition (15b) is identical to (14) implying that the chosen value of L is not only independent of market structure but also independent of liability rule. To compare *consumer welfare* in the different cases it is sufficient to evaluate the total ownership cost H. By use of the definition in (4) one can readily verify that (9) is identical to (15a). It follows that consumer welfare under monopoly is the same for either liability rule. It is obvious that the same consumer welfare prevails under both liability rules when markets are competitive. Since L is the same in all cases, we obtain the standard result that consumer welfare is lower under monopoly than under perfect competition.

While consumer welfare is the same un-

der either liability rule, monopolist's profit will be higher under producer liability. Since we have shown that L and P are the same for both liability rules, the difference in profits is obtained by subtracting (7) from (8):

$$Q[P + \lambda\theta(P + L)B][P - C(L)](\lambda - 1)\theta B$$

When there is loading on insurance, this expression is positive thus verifying that the monopolist's profit is higher under producer liability. This difference arises from the differing amounts of insurance purchased in the two cases. While the consumer insures for the purchase price and the damage $P + L$, the producer insures for the replacement cost and the damage $C + L$. Therefore, the insurance cost is higher under consumer liability, and this additional insurance cost is borne entirely by the producer.

B. *Determination of the Frequency of Failure*

We analyze the frequency of failure and consumer welfare considering, as in part A, the alternative liability rules and market structures.

CASE 1: *Consumer Liability—Monopoly*

The first-order conditions for P and θ are obtained from the objective function (7):

$$\text{(16a)} \quad \frac{\partial \pi}{\partial P} = [Q_1 + (P_1 - C_1)Q_1' \cdot (1 + \lambda\theta_1 B)](1 + \theta_1 B) = 0$$

$$\text{(16b)} \quad \frac{\partial \pi}{\partial \theta} = [-C_1'Q_1 + (P_1 - C_1)Q_1' \cdot \lambda A_1 B] \cdot (1 + \theta_1 B) + (P_1 - C_1) \cdot Q_1 B = 0$$

The subscript 1 indicates the functions are evaluated at the optimally chosen values P_1 and θ_1 for Case 1.

CASE 2: *Consumer Liability—Perfect Competition*

The objective function for this case is the same as (12) except that θ rather than L is the argument of the cost function. The two conditions that determine the optimal values P_2 and θ_2 for the competitive firm are the zero profit condition and that θ_2 maximizes profits when H is treated as a parameter:

$$\text{(17)} \quad P_2 - C_2 = 0$$

$$\text{(18)} \quad \lambda LB + C_2'(1 + \lambda\theta_2 B) + \lambda BC_2 = 0$$

Clearly condition (18) is equivalent to choosing θ to minimize the total cost of owning the asset: $C(1 + \lambda\theta B) + \lambda\theta LB$.

We now compare product safety under the two market structures with consumer liability. Define η to be the price elasticity of demand:

$$\text{(19)} \quad \eta = -\frac{P}{Q}\frac{dQ}{dP} = -\frac{P}{Q}Q'(1 + \lambda\theta B)$$

Combining equation (16a) with (16b), we obtain

$$\text{(20)} \quad C_1'(1 + \lambda\theta_1 B) = \frac{P_1 B}{\eta_1}\left(\frac{1 + \lambda\theta_1 B}{1 + \theta_1 B}\right) - \lambda A_1 B$$

Recalling that $A_1 = P_1 + L$ this is rewritten as

$$\text{(21)} \quad \lambda LB + C_1'(1 + \lambda\theta_1 B) + \lambda BC_1 = \frac{P_1 B}{\eta_1}\left(\frac{1 - \lambda}{1 + \theta_1 B}\right) \leq 0$$

The inequality follows from the fact that $\lambda \geq 1$. Notice that the expression on the left-hand side of (21) is of the same form as (18) which is the first-order condition determining θ_2. Therefore, by the concavity in θ of the function we obtain that $\theta_2 \geq \theta_1$. We thus have shown that under consumer liability with full insurance the monopolist produces an asset with lower probability of failure than the competitive firm. In the special case of actuarially fair insurance $\lambda = 1$, the probability of failure is the same regardless of market structure. Further discussion of the implications of the above results is presented after the analysis of the case of producer liability.

CASE 3: *Producer Liability—Monopoly*

The monopolist's first-order conditions for H_3 and θ_3 obtained from (8) are

$$\text{(22a)} \quad \frac{\partial \pi}{\partial H} = Q_3 + [H_3 - C_3(1 + \lambda\theta_3 B) - \lambda\theta_3 LB]Q_3' = 0$$

$$\text{(22b)} \quad \frac{\partial \pi}{\partial \theta} = -Q_3[C_3'(1 + \lambda\theta_3 B) + C_3\lambda B + \lambda LB] = 0$$

CASE 4: *Producer Liability—Perfect Competition*

For a competitive firm H is taken as given and the equilibrium conditions determining H_4 and θ_4 are zero profit and a condition identical to (22b). Therefore it follows immediately that $\theta_3 = \theta_4$. Moreover, condition (22b) is identical to (18) implying that $\theta_3 = \theta_2$.

In summary, we have shown that $\theta_1 \leq \theta_2 = \theta_3 = \theta_4$. The monopolist's decision to produce a safer product under consumer liability can be understood by considering the difference in his profit between the two liability rules for a given P and θ. By subtracting (8) from (7) we obtain

$$\text{(23)} \quad -Q[P + \lambda\theta(P + L)B](P - C(\theta))(\lambda - 1)\theta B$$

This difference in expected profit represents the additional insurance cost arising under consumer liability due to the fact that the consumer insures against purchase price and the producer insures against production cost. To reduce this extra insurance cost to the consumer, the producer under consumer liability reduces θ and thus produces a safer product than under producer liability. The reason that the frequency of failure is the same under either liability rule for competitive firms is that $P = C$, thus eliminating the excess insurance expenditure.

While the above comparison of product safety is interesting, even more important is the comparison of consumer welfare under the two alternative liability rules. Clearly, under competition consumer welfare is the same under either liability rule. Therefore we proceed to compare the full cost to the consumer of owning the asset in the monopoly case under the two liability rules. We will show that $H_1 \geq H_3$, that is, consumer welfare is higher when the monopolist bears the liability. We first rewrite (16a) by substituting for P in terms of H from equation (4) recalling that $A = P + L$:

$$\text{(24)} \quad Q_1 + [H_1 - C_1(1 + \lambda\theta_1 B) - \lambda\theta_1 LB]Q_1' = 0$$

The form of this expression is the same as (22a). Since we have already shown that $\theta_1 \leq \theta_3$, we can compare H_1 and H_3 by determining how the value of H in (24) changes when θ is increased. Differentiating (24) with respect to θ, treating H as a function of θ, we obtain

$$\text{(25)} \quad \frac{dH_1}{d\theta_1} = \frac{Q_1'[C_1'(1 + \lambda\theta_1 B) + C_1\lambda B + \lambda LB]}{+\, \partial^2\pi/\partial H_1^2} \left[\frac{1 + \lambda\theta B}{1 + \theta B}\right] < 0$$

The denominator is negative by the assumed concavity of the profit function with respect to H. It is apparent from (22b) that $C'(1 + \lambda\theta B) + C\lambda B + \lambda LB$ equals zero at θ_3, and, given the concavity of the objective function in θ, it follows that this expression is negative at θ_1. Since Q_1' is also negative, inequality (25) follows. Since $\theta_1 \leq \theta_3$ we conclude that $H_1 \geq H_3$ thereby verifying that consumer welfare is higher when the monopolist bears the liability.

The comparison of monopolist's profit under the two liability rules is straightforward. Since we have shown in (23) that for any P and θ the profit under producer liability is higher than under consumer liability, it follows that the same relationship holds at the optimally chosen values of P and θ. We have thus shown that for both the consumer and the producer, welfare under monopoly is higher when the producer bears the liability. This result arises

because insurance expenditure is lower when the monopolist bears the liability.

C. *Summary*

In parts A and B we analyzed product safety, consumer welfare, and producer profit under different liability rules. We have found that the size of the loss is the same regardless of market structure or liability rule. Consumer welfare is the same under either liability rule for the same market structure, but, as usual, differs according to market structure. Finally, with monopoly, profit is shown to be higher under producer liability.

When the safety characteristic is the probability of failure, we can summarize the results by the following inequalities: $\theta_1 \leq \theta_2 = \theta_3 = \theta_4$, $H_1 \geq H_3 > H_2 = H_4$, and $\pi_1 \leq \pi_3$. Regarding the safety of the product, we find that the probability of failure is lowest in the case of consumer liability (θ_1) under monopoly. When the monopolist bears the liability, a less safe good (θ_3) is produced. Although safety is lower under producer liability, both consumer welfare and producer profit are higher. The exception is the case of fair insurance in which safety, consumer welfare, and producer profit under monopoly are the same regardless of liability rule. Under perfect competition, safety and consumer welfare are the same for either liability rule, and profit in both cases is zero.

Our results demonstrate, perhaps unexpectedly, that the monopolist produces a good which is at least as safe as that produced by a competitive firm. Furthermore, by making the producer liable, the safety of the good will, if anything, be diminished. The one shortcoming of the analysis in this section is the assumption of full insurance when the consumer is liable. However, as shown in Section I, purchasing full insurance is optimal if there is no loading. Therefore for this case our results hold without qualification. For the case $\lambda > 1$ the results are suggestive, but the assumption of full insurance plays a crucial role. This assumption will be relaxed in the next section.

III. Product Safety with Partial Insurance

In this section we undertake an analysis of the case of partial insurance. From the model in Section I it is apparent that an analysis of partial insurance in a multiperiod model is extremely difficult. Moreover, deriving the market demand function when consumers bear part of the risk appears to be intractable unless one is willing to specify a particular form for the utility function and the distribution of the parameters of this function across consumers. In order to make the problem tractable we adopt three simplifying assumptions. First, we adopt the convention of analyzing the behavior of a representative consumer. Second, we assume a perfectly competitive industry. Third, we confine our attention to a one-period model.

Under competitive conditions the characteristics of the good in equilibrium will be the same as those which would result if the consumer chose the characteristics directly.[2] Therefore, we perform the analysis of utility maximization of a representative consumer-producer.

$$(26)\quad EU = \theta U[W - C(\;) - \lambda\theta I - L + I] + (1 - \theta)U[W - C(\;) - \lambda\theta I] = \theta U(X) + (1 - \theta)U(Y)$$

The arguments X and Y are introduced to simplify the notation. Since we focus on a one-period model, there is no replacement if failure occurs. Thus the only loss incident to failure is the damage L. The other expenditures are the insurance premium $\lambda\theta I$

[2]This may be demonstrated formally. For example, when size of the loss is the safety characteristic, the competitive producer chooses L and P by

$$\text{Max } P - C(L)$$

$$\text{subject to} \quad EU[P, L, I^*(P, L)] = \text{Constant}$$

where EU is the consumer's expected utility as in (26), and $I^*(P, L)$ is defined by $\partial EU/\partial I = 0$. The first-order conditions for this problem together with the additional condition $P = C(L)$ are equivalent to (29).

and the price of the risky asset which under perfect competition is equal to production cost $C(\ )$. As before, we will distinguish between the severity of loss and the probability of failure. Therefore, the argument of the cost function will be L in part A and θ in part B.

In the subsections that follow, we consider only the safety characteristics of the product under alternative liability rules. In the competitive case, the welfare comparison can be made a priori. Partial insurance is optimal when there is loading. Since producer liability implies that the consumer purchases a fully insured product, consumer liability is superior because it effectively relaxes the full insurance constraint. Despite the fact that the welfare comparison is obvious, it is of interest to compare the safety features of the product under alternative liability rules.

A. *Severity of Loss*

We now compare the severity of the loss under producer and consumer liability. When the liability is on the producer, the consumer purchases a fully insured good for a total cost

$$C(L) + \lambda\theta L \tag{27}$$

The producer chooses L to minimize total cost giving rise to the following first-order condition:

$$C'(L) + \lambda\theta = 0 \tag{28}$$

The value of L arising from this expression will now be compared to the value arising under consumer liability. In this case the choice variables in (26) are I and L. The first-order conditions are

$$\theta U'(X)(1 - \lambda\theta) - (1 - \theta)U'(Y)\lambda\theta = 0 \tag{29a}$$

$$-\theta U'(X)(1 + C'(L)) - (1 - \theta)U'(Y)C'(L) = 0 \tag{29b}$$

By direct substitution of $U'(X)$ from (29a) into (29b) we obtain a condition determining L which is identical to (28). Therefore we conclude that the severity of the loss is the same under either liability rule.[3]

B. *Frequency of Failure*

Under producer liability, the total cost of a fully insured product is $C(\theta) + \lambda\theta L$. The producer chooses θ to satisfy

$$C'(\theta) + \lambda L = 0 \tag{30}$$

Under consumer liability the optimal I and θ are determined by

$$\theta U'(X)(1 - \lambda\theta) - (1 - \theta)U'(Y)\lambda\theta = 0 \tag{31a}$$

$$U(X) - U(Y) - [\theta U'(X) + (1 - \theta)U'(Y)][C'(\theta) + \lambda I] = 0 \tag{31b}$$

First we will show that when $\lambda > 1$, the optimally chosen insurance level I is less than the size of the loss. By definition of X and Y it is clear that $X = Y - (L - I)$. Therefore it is sufficient to show that in (31a), $X < Y$. From (31a),

$$\frac{U'(X)}{U'(Y)} = \frac{(1 - \theta)\lambda}{1 - \lambda\theta}$$

Since U is concave we must show that the right-hand side is greater than 1. From $\lambda > 1$ it follows that $\lambda(1 - \theta) > 1 - \lambda\theta$. However, $1 - \lambda\theta$ is positive; otherwise even if the loss occurs, the payment from the insurance policy is lower than the premium paid. Hence, if insurance is purchased, $\lambda(1 - \theta)/(1 - \lambda\theta) > 1$ and it follows that $L > I$.

We now compare the frequency of failure under the two liability rules. Substituting from (31a) into (31b) and solving we obtain

$$C'(\theta) + \lambda L = \lambda\frac{U(X) - U(Y)}{U'(X)} + \lambda(L - I) \tag{32}$$

By definition $X = Y - (L - I)$ and by Taylor's expansion, $U(X) - U(Y) = -(L - I)U'(Z)$ where $X < Z < Y$. Therefore (32) can be written as

[3]This result corresponds to Isaac Ehrlich and Gary Becker's demonstration that the degree of self-insurance is independent of the parameters of the utility function.

$$(33)\quad C'(\theta) + \lambda L = \lambda(L - I)\left[1 - \frac{U'(Z)}{U'(X)}\right]$$

Since $Z > X$ and U' is decreasing, $U'(Z) < U'(X)$. Also it was shown above that $L > I$. It follows that the right-hand side of (33) is positive. The value of θ which satisfies (33) must therefore be greater than the value which satisfies (30). Thus we have proved that a less safe product will be optimal under consumer liability than under producer liability if insurance is imperfect.

Intuitively, the result in the present case differs from that in the previous subsection because the change in θ changes the "price" of insurance by changing the premium required per dollar of protection. Therefore, the greater the amount of insurance purchased, the greater the incentive to reduce θ. Since producer liability is equivalent to full insurance, the optimal θ in this case is lower than under consumer liability.

The discussion of the results in parts A and B will follow the analysis of imperfect information developed in the next section.

IV. Product Safety with Partial Insurance and Imperfect Information

The most important assumption implicit in our analysis thus far is that the consumer has perfect information about the safety characteristics of the product. Clearly the consumer rarely has such complete information. It is, therefore, important to determine whether our results will change if we relax this assumption.

In what follows we assume that neither the consumer nor the insurer has perfect information about the safety characteristics of the product. Each of them has a subjective probability density function, not necessarily the same, about the severity of loss and the frequency of failure. The density function for the insurer is $g(l, \varphi)$. Thus the insurance premium for coverage in the amount I is $\lambda I \iint g(l, \varphi)\varphi\, d\varphi\, dl$, which is the expected payment to the insured multiplied by the loading λ. The insurer will generally have more information about the product safety characteristics than the typical consumer, and therefore we assume that his estimate of θ is unbiased. It follows that the premium equals $\lambda\theta I$.

We assume that the producer has full knowledge of the product characteristics. Therefore, under producer liability, equations (28) and (30) determine the optimal values of L and θ, respectively.

The optimal product characteristics in a competitive industry under consumer liability are obtained from maximizing the expected utility of a representative consumer-producer,

$$(34)\quad EU = \iint f(l, \varphi)\{\varphi U[W - C(\ ) - \lambda\theta I - l + I] + (1 - \varphi)U[W - C(\ ) - \lambda\theta I]\}\, d\varphi\, dl$$

with respect to I and L or θ. The consumer's perceptions regarding product safety characteristics are represented by a subjective probability density function $f(l, \varphi)$. In general, this function depends on the true values of L and θ. In the sections that follow this dependence will be made explicit. To highlight the effects of uncertainty about each dimension of safety, we will analyze the effects of partial information regarding one of the safety characteristics while assuming that the other is known with certainty. Thus the density function will be either $f(l)$ or $f(\varphi)$, and the argument of the cost function will be L and θ in parts A and B, respectively.

A. *Severity of Loss*

When the liability is on the producer, the condition determining L is equation (28). Under consumer liability the choice variables in (34) are I and L. In order to derive qualitative results the functional dependence of $f(l)$ on L must be specified. We will consider two alternative specifications which seem to us to capture the essence of the effects of imperfect information. In the first case we assume that the perceived size of the loss is the true but unknown loss plus a random noise. This implies that the degree of uncertainty regarding the size of the loss is independent of L. In the second case we assume that the perceived size of the loss is

a random variable proportional to the true loss. In this case the degree of uncertainty increases with the size of the true loss.

The first assumption can be stated formally as $l = L + z$ where z is a random variable whose distribution is not functionally dependent on L. With this change of variable and an obvious redefinition of the density function f, equation (34) becomes

$$(35)\quad EU = \theta \int f(z) U[W - C(\ ) - \lambda\theta I - L - z + I]\, dz + (1 - \theta) U[W - C(\ ) - \lambda\theta I]$$

where $f(z)$ is not functionally dependent on L.

The first-order conditions are:

$$(36a)\quad (1 - \lambda\theta)\theta \int f(z) U'(X)\, dz - \lambda\theta(1 - \theta) U'(Y) = 0$$

$$(36b)\quad -(1 + C'(L))\theta \int f(z) U'(X)\, dz - C'(L)(1 - \theta) U'(Y) = 0$$

By direct substitution we obtain that L is determined by a condition identical to (28). Therefore, we conclude that even though information is imperfect, the size of the loss will be the same under either liability rule. This result obtains even if the consumer's expectations are biased.

While the size of the loss does not depend on the information available to the consumer or the liability rule, both of these will affect consumer welfare. There is a certain degree of ambiguity in defining welfare in this case. If the consumer underestimates the probability of failure or size of the loss, his subjective expected utility may be higher than under the perfect information case. In our view, higher or lower subjective levels of utility due to misperception should not be counted as improving or diminishing welfare. Rather, consumer welfare should be defined as the expected utility achieved when the actual insurance purchasing decision is evaluated with the correct probability of failure and size of loss. Since the actual insurance purchased is determined from (36a), in general, it will be different from that purchased under perfect information. The objective welfare level will, therefore, generally be lower under consumer liability when information is imperfect.

In contrast to the perfect information case, the objectively determined welfare level under consumer liability may now be lower than the welfare level under producer liability. Producer liability is equivalent to purchase of full insurance coverage. When the consumer has imperfect information, he might purchase coverage for an amount that exceeds the actual loss if failure were to occur.[4] A demonstration that this might occur is provided in the following example.

For this example we assume that the consumer estimates L without bias. It follows that $E(z) = 0$. Moreover, we introduce a parameter k describing the "dispersion" of the subjective distribution of z. Rewrite (35) as

$$(37)\quad EU = \theta \int f(z) U[W - C(L) - \lambda\theta I - L - zk + I]\, dz + (1 - \theta) U[W - C(L) - \lambda\theta I]$$

An increase in k can be thought of as a mean preserving increase in the perceived risk. This increase in perceived risk is equivalent to a reduction in the consumer's information about the safety characteristics of the product. Since it was already proved that L is independent of the perceived risk, $I(k)$ is obtained for any k from

$$\frac{\partial EU}{\partial I} = \theta(1 - \lambda\theta) \cdot \int f(z) U'[W - C(L) - \lambda\theta I - L - zk + I]\, dz - \lambda\theta(1 - \theta) U'[W - C(L) - \lambda\theta I] = 0$$

The comparative static response of $I(k)$ to changes in k is obtained by differentiating the above function

[4]This excessive insurance purchase is not prevented by the insurer since his information is imperfect as well. If the insurer had perfect information it is clear that the purchase of insurance would be restricted and welfare under consumer liability could not be lower than under producer liability.

$$\frac{dI}{dk} = \theta(1 - \lambda\theta) \int z f(z) U'' [W - C(L) - \lambda\theta I - L - zk + I] dz + \partial^2 EU/\partial I^2$$

The denominator is negative by the second-order conditions. Using integration by parts, the numerator can be shown to be negative if $U''' > 0$ which is the case if decreasing absolute risk aversion is assumed. Since the level of insurance increases with increased risk, it is quite possible that a risk averter would purchase insurance above L when his information is below some critical level. If this were to occur, then welfare under consumer liability would be lower than under producer liability.

In the second case we assume a multiplicative random variable $l = zL$ where z is a random variable taking on positive values with density $f(z)$ not functionally dependent on L. We also assume that the consumer's perception is unbiased, i.e., $E(l) = L$, which implies that $E(z) = 1$. We can now rewrite (34) as

$$\text{(38)} \quad EU = \theta \int f(z) U[W - C(L) - \lambda\theta I - zL + I] dz + (1 - \theta) U[W - C(L) - \lambda\theta I]$$

The first-order conditions for I and L are

$$\text{(39a)} \quad \theta(1 - \lambda\theta) \int f(z) U'(X) dz - \lambda\theta(1 - \theta) U'(Y) = 0$$

$$\text{(39b)} \quad -\theta \int [z + C'(L)] f(z) U'(X) dz - (1 - \theta) C'(L) U'(Y) = 0$$

By direct substitution, the above conditions can be rewritten as

$$\text{(40)} \quad \int [C'(L) + \lambda\theta z] f(z) U'(X) dz = 0$$

We now demonstrate that the size of the loss satisfying (40) is lower than that obtained under producer liability. Under producer liability, the size of the loss L_1 determined in (28) satisfies

$$\text{(41)} \quad \int [C'(L_1) + \lambda\theta] f(z) U'(X) dz = 0$$

Together, (40) and (41) imply

$$\text{(42)} \quad \int \lambda\theta(z - 1) f(z) U'(X) dz = \int [C'(L_1) - C'(L)] f(z) U'(X) dz$$

Integrating by parts, the left-hand side equals

$$L \int_0^\infty U''(X) \left\{ \int_0^s (z - 1) f(z) dz \right\} ds$$

which is positive since $U'' < 0$ and

$$\int_0^s (z - 1) f(z) dz < E(z - 1) = 0$$

Therefore, it follows from the right-hand side of (42) that $C'(L_1) > C'(L)$. Since from (6) $C'' > 0$, we verify our claim that $L_1 > L$. Thus, when the degree of uncertainty increases with the size of the true loss, a safer product is chosen under consumer liability. As in the previous case, the welfare comparison depends on the amount of information available to consumers and the extent of their risk aversion. A risk-averse consumer might purchase excessive insurance to the point where the total cost of purchasing and insuring the product is higher under consumer liability than under producer liability.

B. *Frequency of Failure*

When the liability is on the producer, the condition determining θ is equation (30). Since we have already investigated the effects of lack of information about the size of the loss in this section we simplify by assuming that there is no uncertainty regarding L. Then (34) is rewritten:

$$\text{(43)} \quad EU = \hat{\theta} U[W - C(\theta) - \lambda\theta I - L + I] + (1 - \hat{\theta}) U[W - C(\theta) - \lambda\theta I]$$

where $\hat{\theta}$ denotes the expected value of φ. It is apparent from (43) that this case will not differ from the perfect information case if the consumer's expectations are unbiased. To investigate the effect of biased expectations the first-order conditions for I and θ are derived.

$$(44a)\qquad \frac{\partial EU}{\partial I} = (1 - \lambda\theta)\hat{\theta}U'(X) - \lambda\theta(1 - \hat{\theta})U'(Y) = 0$$

$$(44b)\qquad \frac{\partial EU}{\partial \theta} = \frac{d\hat{\theta}}{d\theta}[U(X) - U(Y)] - [\hat{\theta}U'(X) + (1 - \hat{\theta})U'(Y)][C'(\theta) + \lambda I] = 0$$

We will first show how the insurance decision is affected by variations in $\hat{\theta}$. From (44a),

$$(45)\qquad \frac{U'(X)}{U'(Y)} = \frac{\lambda\theta(1 - \hat{\theta})}{(1 - \lambda\theta)\hat{\theta}}$$

If $\hat{\theta} \geq \lambda\theta$, the right-hand side of (45) is less than or equal to 1. Therefore, for this case (45) would imply that $X \geq Y$, i.e., $I \geq L$. Intuitively, the consumer perceives insurance to be a "bargain," that is, the premium is perceived to be less than the expected payment from insurance. Therefore he chooses to either fully insure or to over-insure thus engaging in what is perceived as better than a fair bet. However, we have assumed that the consumer knows L with certainty, and we will assume that the insurer knows the true potential loss with certainty as well, and that he will impose the constraint that the insurance cannot exceed the potential loss. The appropriate first-order conditions would then imply that (45) holds as an inequality. Thus $\hat{\theta} \geq \lambda\theta$ implies $I = L$, that is, full insurance, which together with (44b) yields the condition determining θ which is equivalent to (30). Thus when $\hat{\theta} \geq \lambda\theta$ under consumer liability, we obtain the same frequency of failure and the same level of consumer welfare as in the producer case.

We now investigate the case $\hat{\theta} < \lambda\theta$. In this case insurance will be less than full and the consumer's misperception will induce the producer to change the characteristic of the product. To illustrate this effect, we consider the following example in which the mean of the subjective probability distribution of frequency of failure is proportional to the true frequency of failure, i.e., $\hat{\theta} = m\theta$. The estimated mean is biased upward or downward as m is greater or less than one. For the case under consideration $m < \lambda$.

We first investigate how changes in the bias in the consumer's perception, that is, changes in m affect the true probability of failure. Clearly, where there is no bias, $m = 1$, the first-order conditions (44) are identical to those under perfect information, equations (31). To determine the effect of a change in m, we perform a comparative static analysis of equations (44):

$$\frac{d\theta}{dm} = -\frac{\theta U'(Y)}{|D|}\{(1 - \lambda\theta)U''(X)[C'(\theta) + \lambda I] + \lambda^2\theta[U'(X) - U'(Y)]\}$$

where $|D|$ is the determinant of the second-order derivatives. From (44b) we observe that $C'(\theta) + \lambda I$ is negative; $|D|$ is positive by the second-order conditions. It follows by inspection that $d\theta/dm < 0$. Thus, the true safety of the product changes in a direction opposite the direction of change in the bias. By inducing a change in the built-in frequency of failure, the consumer's misperception leads to a change in the objective welfare level. In general, one would expect the objectively measured level to decrease as the bias increases. However, welfare under consumer liability cannot be below that of producer liability since, in this case, the consumer cannot buy more than full insurance.

C. *Summary*

Contrasting the results of Section IIIA with those in IVA we note that if consumer's information is imperfect, the size of the loss may differ according to the liability rule. In addition, welfare under consumer liability is reduced by lack of information, and it is possible that this reduction will make welfare under consumer liability lower than under producer liability.

The effect of imperfect information on the frequency of failure can be seen by comparing the results in Sections IIIB and IVB. Under consumer liability, the frequency of

failure is unaffected by the dispersion of the subjective distribution as long as the mean of this distribution equals the true value of θ. In this case a less safe product will be produced under consumer liability. When bias in the subjective beliefs is introduced, we find that the safety changes in a direction opposite to the change in the bias. Therefore, if the consumer underestimates the probability of failure, then the true probability of failure will be higher than under consumer liability with perfect information, which in turn is higher than under producer liability. Lack of information reduces welfare under consumer liability, but the welfare cannot be lower than under producer liability.

V. Conclusions

Our results indicate the crucial role of insurance and information in the determination of product characteristics and consumer welfare. When insurance is perfect, that is, actuarially fair, and consumers have perfect information, product safety characteristics were shown to be independent of market structure and liability rules. When either insurance or information is imperfect, these results do not necessarily hold.

We find that no single liability rule is universally applicable if there are imperfections in either the insurance market or in consumer information. If only full insurance is available and information is perfect our results in Section II indicate that the assignment of liability will not generally matter. The one exception occurs under monopoly where producer liability results in a product which fails more frequently. This exception is only important if the replacement of the good constitutes a significant portion of the total loss when failure occurs.

If partial insurance can be purchased and information is perfect, then consumer liability yields higher welfare. This is true even though the probability of failure in this case is higher than under producer liability.

When information is imperfect, product safety characteristics and consumer welfare depend on both the degree of uncertainty and the extent of the bias in consumer perceptions. The optimal liability assignment also depends on consumer information as well as his risk preferences. As a general rule, consumer liability is preferable for products for which the consumer can judge safety characteristics with reasonable accuracy. On the other hand, producer liability may be desirable if it is difficult for the consumer to evaluate the product. The National Commission on Product Safety argued that products should be classified either as "reasonably" or "unreasonably" hazardous depending on the consumer's knowledge of the risks associated with use of the product.[5] The analysis in this paper shows that such a classification may be useful in determining the optimal liability rule.

We believe that the analysis developed in this paper has significantly improved our understanding of the problems related to product safety. Our results apply to the class of products for which the safety characteristics are determined by the manufacturer and thereafter cannot be modified by the consumer. The other class of products are those for which the occurrence of accidents is affected not only by the characteristics determined at the time of manufacture, but also by the care exercised by the consumer in the use of the product. This class of products was beyond the scope of this paper and remains an important subject for future research.

Appendix

In this Appendix we maximize (2) subject to the wealth equations (3). The solution is via stochastic dynamic programming. Denoting by $V_t(W_t, I)$ the optimal level of utility from time t and onward when the wealth level at t is W_t and the level of insur-

[5]Oi, p. 4, reproduces the definitions endorsed by the National Commission on Product Safety.

ance is I, we obtain the following recursive equations:

(A1) $$V_T(W_T, I) = U(W_T)$$

(A2) $$V_{t-1}(W_{t-1}, I) = \underset{C_{t-1}}{\text{Max}} \Big\{ U(C_{t-1}) + \gamma\theta V_t\left[\frac{1}{\beta}(W_{t-1} - C_{t-1}) - A + I, I\right] + \gamma(1-\theta)V_t\left[\frac{1}{\beta}(W_{t-1} - C_{t-1}), I\right]\Big\}$$

for t = 2, 3, ..., T

(A3) $$V_0(W_0) = \underset{C_0, I}{\text{Max}} \Big\{ U(C_0) + \gamma\theta V_1\left[\frac{1}{\beta}(W_0 - C_0 - R) - A + I, I\right] + \gamma(1-\theta)V_1\left[\frac{1}{\beta}(W_0 - C_0 - R), I\right]\Big\}$$

Equation (A1) simply states that all wealth remaining at the beginning of period T is consumed. At the beginning of periods 1 through T − 1, consumption is chosen as indicated by (A2) to maximize discounted expected utility. At the beginning of period 0 in addition to the consumption decision, the consumer chooses the level of insurance coverage which will be maintained in all subsequent periods. The premium R is related to the insurance coverage I through equation (1).

The first-order conditions for the consumption decisions are

(A4) $$U'(C_{t-1}) - \frac{\gamma}{\beta}\theta V'_t\left[\frac{1}{\beta}(W_{t-1} - C_{t-1}) - A + I, I\right] - \frac{\gamma}{\beta}(1-\theta)V'_t\left[\frac{1}{\beta}(W_{t-1} - C_{t-1}), I\right] = 0$$

for t = 2, 3, ..., T

(A5) $$U'(C_0) - \frac{\gamma}{\beta}\theta V'_1\left[\frac{1}{\beta}(W_0 - C_0 - R) - A + I, I\right] - \frac{\gamma}{\beta}(1-\theta)V'_1\left[\frac{1}{\beta}(W_0 - C_0 - R), I\right] = 0$$

In the above equations V'_t denotes differentiation with respect to the first argument. The first-order condition determining I is

(A6) $$-\frac{dR}{dI}\Big\{\frac{\gamma}{\beta}\theta V'_1\left[\frac{1}{\beta}(W_0 - C_0 - R) - A + I, I\right] + \frac{\gamma}{\beta}(1-\theta)V'_1\left[\frac{1}{\beta}(W_0 - C_0 - R), I\right]\Big\} + \gamma\theta V'_1\left[\frac{1}{\beta}(W_0 - C_0 - R) - A + I, I\right] + \gamma\theta\frac{\partial V_1}{\partial I}\left[\frac{1}{\beta}(W_0 - C_0 - R) - A + I, I\right] + \gamma(1-\theta)\frac{\partial V_1}{\partial I}\left[\frac{1}{\beta}(W_0 - C_0 - R), I\right] = 0$$

We now investigate whether purchasing full insurance, i.e., $I = A$, is the optimal decision. This is done by analyzing the first-order conditions (A4) and (A5) at $I = A$. When $I = A$ the occurrence of the loss does not change the wealth level as can be seen from (3). We will denote by W_t^* and C_t^* the wealth and consumption levels at time t when consumption is optimally chosen in all previous periods. Thus $W_t^* = (1/\beta)(W_{t-1}^* - C_{t-1}^*)$ for t = 2, 3, ..., T and $W_1^* = (1/\beta)(W_0 - C_0^* - R)$. At $I = A$ (A4) and (A5) become

(A7a) $$U'(C_{t-1}^*) - \frac{\gamma}{\beta}V'_t(W_t^*, I) = 0$$

for t = 2, 3, ..., T

(A7b) $$U'(C_0^*) - \frac{\gamma}{\beta}V'_1(W_1^*, I) = 0$$

Equation (A6) becomes

(A8) $$-\frac{dR}{DI}\frac{\gamma}{\beta}V'_1(W_1^*, I) + \gamma\theta V'_1(W_1^*, I) + \gamma\frac{\partial V_1}{\partial I}(W_1^*, I) = 0$$

We will now derive $\partial V_1/\partial I$. Differentiating (A2) with respect to I and using the first-

order condition (A7a) for C_{t-1}, we get

$$\text{(A9)}\quad \frac{\partial V_{t-1}(W^*_{t-1},I)}{\partial I} = \gamma\theta V'_t(W^*_t,I) + \frac{\gamma\partial V_t(W^*_t,I)}{\partial I} \quad \text{for } t = 2,3,\ldots,T$$

By successive lagging and substitution, equation (A9) is rewritten as

$$\text{(A10)}\quad \frac{\partial V_1(W^*_1,I)}{\partial I} = \theta \sum_{t=2}^{T} \gamma^{t-1} V'_t(W^*_t,I)$$

Differentiating (A2) with respect to W_{t-1}, we obtain

$$\text{(A11)}\quad V'_{t-1}(W^*_{t-1},I) = \frac{\gamma}{\beta} V'_t(W^*_t,I) \quad \text{for } t = 2,3,\ldots,T$$

which implies that

$$\text{(A12)}\quad V'_t(W^*_t,I) = \left(\frac{\beta}{\gamma}\right)^{t-1} V'_1(W^*_1,I) \quad \text{for } t = 2,3,\ldots,T$$

Substituting (A12) into (A10),

$$\text{(A13)}\quad \frac{\partial V_1(W^*_1,I)}{\partial I} = \theta V'_1(W^*_1,I) \sum_{t=2}^{T} \beta^{t-1}$$

From equation (1),

$$\text{(A14)}\quad \frac{dR}{dI} = \lambda\theta \sum_{t=1}^{T} \beta^t$$

By substituting (A13) and (A14) into (A8) we immediately verify that the first-order condition is satisfied at $I = A$ if and only if $\lambda = 1$. Thus we have proved that it is optimal to purchase full coverage if and only if insurance is actuarially fair.

When $\beta = \gamma$ we obtain from (A12) that $V'_t(W^*_t,I) = V'_1(W^*_1,I)$ for t = 2, 3, ..., T. Using this in the first-order conditions (A7), it follows immediately that $C^*_0 = C^*_1 = \ldots = C^*_T$ as asserted in the text.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1970.

Richard E. Barlow, and Frank Proschan, *Mathematical Theory of Reliability*, New York 1965.

J. P. Brown, "Product Liability: The Case of an Asset with Random Life," *Amer. Econ. Rev.*, Mar. 1974, *64*, 149-61.

J. Buchanan, "In Defense of *Caveat Emptor*," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 64-73.

G. Calabresi, and K. C. Bass III, "Right Approach, Wrong Implications: A Critique of McKean on Products Liability," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 74-91.

R. Dorfman, "The Economics of Products Liability: A Reaction to McKean," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 92-102.

I. Ehrlich and G. S. Becker, "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*, 623-48.

V. Goldberg, "The Economics of Product Safety and Imperfect Information," *Bell J. Econ.*, Autumn 1974, *2*, 683-88.

K. Hamada, "Liability Rules and Income Distribution in Product Liability," *Amer. Econ. Rev.*, Mar. 1976, *66*, 228-34.

R. N. McKean, "Products Liability: Trends and Implications," *Univ. Chicago Law Rev.*, Fall 1970, *38*, 3-63.

———, "Products Liability: Implications of Some Changing Property Rights," *Quart. J. Econ.*, Nov. 1970, *84*, 611-26.

W. Y. Oi, "The Economic Analysis of Product Safety," *Bell J. Econ.*, Spring 1973, *4*, 3-28.

Energy, the Heckscher-Ohlin Theorem, and U.S. International Trade

By ARYE L. HILLMAN AND CLARK W. BULLARD III*

The Heckscher-Ohlin model has lost credibility as a positive instrument of analysis. The prime cause is its persistent indication of a direction of *U.S.* comparative advantage in international trade which is the converse of that which informed observation suggests ought to result from a relative capital-labor factor proportions test of *U.S.* data. In his pioneering studies, Wassily Leontief (1953, 1956) using the 1947 input-output structure with both 1947 and 1951 trade data found exports relatively labor intensive, and import-competing production relatively capital intensive. Robert Baldwin (1971) employing the 1958 input-output structure and 1962 trade data reaffirmed the Leontief paradox.[1] In this paper we continue in the tradition of the Leontief-Baldwin studies by investigating the Heckscher-Ohlin Theorem and *U.S.* trade on the basis of the 1963 and 1967 input-output production structures and concurrent trade flows; our methods differ from the previous studies in that we employ energy as a reference factor of production. Although strictly speaking energy is a produced factor of production, we do not treat it as an intermediate good. Rather we view it as a nonproduced input symmetric with labor and capital.[2] As such, we focus on the raw material content of energy inputs. The treatment of energy as a nonproduced factor yields a model with three inputs and two traded outputs. Such a model offers a more flexible interpretation of the relation between relative factor endowments and intensities than the traditional two-input, two-output setting in which the services of capital and labor must necessarily be exchanged in different directional trade flows. We further employ as a working hypothesis the supposition that energy and capital exhibit technological complementarity. Support for this supposition derives from the estimation by Ernst Berndt and David Wood of an energy-inclusive cost function for the United States. In the estimated transcendental logarithmic function, Berndt and Wood were unable to reject the conditions for energy-capital linear separability, and hence were unable to reject the conditions for consistent aggregation of energy and capital. On the basis of complementarity we form a Hicksian composite input from energy and capital and thereby proceed without independent capital data by employing energy as a surrogate for capital. By this procedure we circumvent the traditional difficulties associated with the interpretation and aggregation of capital data. After the formulation of the model and suggestion of hypotheses in Section I, we present our empirical results in Section II, with amendments in Section III to accommodate possible special influences due

*Department of economics, Tel-Aviv University, and Center for Advanced Computation, University of Illinois-Urbana, respectively. We are grateful to George Borts and Robert Baldwin for very helpful guidance. We have also benefited from the comments of an anonymous referee. The support of the National Science Foundation and the Foerder Institute for Economic Research, Tel-Aviv University, is acknowledged.

[1]The failure of the Heckscher-Ohlin model to yield the direction of comparative advantage predicted from the assumption of *U.S.* abundance of capital relative to labor has prompted a vast literature offering various explanations of the Leontief paradox. The general consensus from recent studies of the determinants of comparative advantage is that the Heckscher-Ohlin model is too restrictive in the explanatory variables which it offers: see for example Baldwin (1971) and Robert Stern. However, the Heckscher-Ohlin model remains the principal normative setting for the general equilibrium theory of international trade.

[2]One may wish to claim that capital is also a produced input. See, for example, Joan Robinson for a discussion relevant to the issues this raises. On the special characteristics of energy as a factor of production which render it physically nonsymmetric with other inputs, see Nicholas Georgescu-Roegen.

to natural resource sectors. The conclusions are presented in Section IV.

I. The Model

We consider a competitive, externality-free, linear Leontief economy[3] with given factor endowments $\bar{E}$, $\bar{K}$, and $\bar{L}$ of energy, capital, and labor. The economy produces two final consumption traded goods A and B, employing the fixed coefficient constant-returns-to-scale technology

$$(1)\quad \begin{bmatrix} A \\ B \end{bmatrix} = \min \Lambda \cdot (E, K, L, R_1 \ldots R_n)'$$

where Λ is the Leontief matrix of output-input coefficients and $(R_1 \ldots R_n)$ are non-traded intermediate goods. Eliminating the latter from (1) yields

$$(2)\quad \begin{bmatrix} A \\ B \end{bmatrix} = \min \psi \cdot (E, K, L)'$$

where the elements of ψ are direct plus indirect output-input coefficients. Energy, capital, and labor are intersectorally homogeneous and mobile. Homogeneous capital has the same attendant complementary energy requirement independent of its sectoral employment. Accordingly,

$$(3)\quad a_K/a_E = b_K/b_E \equiv \gamma$$

where a and b are designated input-output coefficients. Hence energy and capital constitute a Hicksian composite input.

Domestic competition yields the price-cost relations

$$(4)\quad \begin{bmatrix} P_A^0 \\ \\ P_B^0 \end{bmatrix} \leq \begin{bmatrix} a_E & a_K & a_L \\ \\ b_E & b_K & b_L \end{bmatrix} \begin{bmatrix} e^0 \\ r^0 \\ w^0 \end{bmatrix}$$

FIGURE 1

where 0 denotes a domestic value in the price vector (P_A, P_B, e, r, w). The dual Rybczynski equations are

$$(5)\quad \begin{bmatrix} a_E & b_E \\ a_K & b_K \\ a_L & b_L \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \bar{E} \\ \bar{K} \\ \bar{L} \end{bmatrix}$$

As a consequence of (3), the first and second rows of the coefficient matrix in (5) are linearly dependent. We assume that the domestic Scitovsky preference map is homothetic and that the economy's transactions with the rest of the world are in balance.

Let the economy's factor endowments be such as to yield a surplus of capital, and correspondingly, a deficiency of energy, relative to the technologically given composition of the Hicksian input, so $\bar{E}/\bar{K} < \gamma$.[4] We further assume for reference that good A is relatively labor intensive. The implied Rybczynski lines are depicted in Figure 1.

Consider now an autarkic equilibrium. The equilibrium shadow price of capital is

[3]Our assumption of a Leontief economy is, of course, not a reflection of a belief concerning actual production specification, but provides a consistent frame of reference for the use of input-output studies and data.

[4]If capital is produced, this raises the question why was it produced in excess of the availability of complementary energy. If we enter into this question, we have a third produced good. To focus on our prime consideration, we simply treat capital as a historical endowment.

zero, which is attained by competitive supply bidding of owners of domestic capital. Suppose the autarkic production-consumption equilibrium to obtain at the intersection G of the binding energy and labor constraints. The economy is then diversified in production, and with P_A^0 and P_B^0 given by domestic demand conditions, the positive autarkic price of energy and positive autarkic wage are derived from (4). Energy secures the entire return accruing to the composite input.

Permitting the economy to trade in A and B at equilibrium world prices P_A^* and P_B^* and employing the Heckscher-Ohlin assumption of internationally identical production technologies yields the competitive price-cost relations

$$(6)\quad \begin{bmatrix} P_A^* \\ P_B^* \end{bmatrix} = \begin{bmatrix} a_E & a_K & a_L \\ b_E & b_K & b_L \end{bmatrix} \begin{bmatrix} e^* \\ r^* \\ w^* \end{bmatrix} \leqq \begin{bmatrix} a_E & a_K & a_L \\ b_E & b_K & b_L \end{bmatrix} \begin{bmatrix} e^0 \\ r^0 \\ w^0 \end{bmatrix}$$

Assume $a_E/b_E < P_A^*/P_B^* < a_L/b_L$, so the production equilibrium in free trade remains at G, and let the foreign factor prices e^*, r^*, w^* be strictly positive. Since the free trade production equilibrium is diversified, (6) holds with equality. The equilibrium return to capital remains zero, and substituting $r^0 = 0$ into (6) we have

$$(7)\quad \phi \begin{bmatrix} e^0 \\ w^0 \end{bmatrix} = \phi \begin{bmatrix} e^* \\ w^* \end{bmatrix} + r^* \begin{bmatrix} a_K \\ b_K \end{bmatrix}$$

where $\phi = \begin{bmatrix} a_E & a_L \\ b_E & b_L \end{bmatrix}$

From (3),

$$(8)\quad \begin{bmatrix} a_K \\ b_K \end{bmatrix} = \gamma \begin{bmatrix} a_E \\ b_E \end{bmatrix} = \gamma \phi_1$$

where ϕ_1 denotes the first column of the coefficient matrix ϕ. The factor intensities of A and B differ and hence the inverse ϕ^{-1} exists. Employing (8) and the relation then that

$$(9)\quad \phi^{-1}\phi_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

permits (7) to be solved for the international factor-price equality,

$$(10)\quad \begin{bmatrix} e^0 \\ w^0 \end{bmatrix} = \begin{bmatrix} e^* \\ w^* \end{bmatrix} + \begin{bmatrix} \gamma r^* \\ 0 \end{bmatrix}$$

Hence, in the free-trade equilibrium,

$$(11)\quad w^0 = w^*$$

$$(12)\quad e^0 = e^* + \gamma r^* > e^*$$

$$(13)\quad r^0 = 0 < r^*$$

Factor-price equalization obtains for labor, and thereby for the composite input, but the composition of the return to the latter input differs internationally.

There still now exist potential gains from the international relocation of factors which have not been embodied in the indirect factor content of trade in goods. Accordingly, let us admit international factor mobility. Equation (11) indicates that labor is in equilibrium with no incentive to move. Equations (12) and (13), however, respectively indicate an incentive to import energy directly and to export capital directly. We perceive of capital mobility, rather than entailing the international transfer of claims over productive assets (see George Borts), as a physical transfer which involves a change in location, but not ownership, of real capital equipment.[5] In a free trade post-factor movement equilibrium, the economy's capital and energy Rybczynski lines coincide, and

$$(14)\quad \begin{bmatrix} e^0 \\ r^0 \end{bmatrix} = \begin{bmatrix} e^* \\ r^* \end{bmatrix}$$

Complete factor-price equalization accordingly obtains. That is, in addition to the equalization induced by free trade in the returns to labor and the composite input,

[5]See Ronald Jones, p. 3. See also Baldwin (1966) and Murray Kemp.

we further obtain equalization in the composition of the components of the composite input's return.

In principle, the postfactor movement equilibrium could be established by an outflow only of capital, an inflow only of energy, or a combination of capital exports and energy imports. Suppose that international movements of both capital and energy take place such that the postfactor movement feasible production frontier is observed as *DJS*. Since the world terms of trade are independent of factor movements, *J* is the point of domestic production equilibrium. At *J*, capital is exported directly and energy is imported directly: we are interested in ascertaining the attendant nature of indirect commodity-embodied factor trade, and in comparing the comparative advantage outcome with the predictions of the Heckscher-Ohlin Theorem.

An application of the traditional statement of the Heckscher-Ohlin Theorem to an investigation of international trade at a point such as *J* is compromised by two considerations. We have three inputs and two outputs; and further, the theorem generally assumes factors to be immobile, so that factor trade occurs only indirectly. We state the theorem, following Jaroslav Vanek (1968), as offering the following prediction:

For a given ranking of domestic-foreign relative factor endowments, $X_1^0/X_1^* \geqq X_2^0/X_2^* \geqq \ldots \geqq X_n^0/X_n^*$, *the domestically designated economy will be a net exporter of the services of the factors* $X_1, X_2, \ldots, X_j$ *and will be a net importer of the services of* $X_{j+1}, X_{j+2}, \ldots, X_n$, $j \neq n$.

Here, net factor trade is inclusive of commodity-embodied and direct factor flows. The theorem as stated requires as a necessary condition factor-price equalization,[6] which we have seen to obtain at *J*.

The model is closed with the introduction of demand. We assume that demand reversals of the pattern of comparative advantage do not occur. Figure 1 contains three income-expansion paths which could in principle obtain from the domestic homothetic preference map at the equilibrium world terms of trade. Preferences as would underly the path (1a) indicate a comparative advantage at *J* in the good *B*, which is relatively intensive in the composite input. The same pattern of comparative advantage would obtain in the absence of factor movements in a free trade equilibrium at *G*, where the volume of trade would be smaller. Preferences giving rise to a path (1b), which passes between *G* and *J*, also yield a comparative advantage in *B* at the equilibrium *J*; however, factor mobility will have reversed the pattern of comparative advantage from that obtaining at *G*. An observation that *B* is exported at *J* does not permit the data to yield a distinction between (1a) and (1b). The expansion path (2) leads the economy to exhibit a comparative advantage at *J* in *A*, which is relatively labor intensive. At *G* the same comparative advantage would be exhibited and hence in this case the direction of trade is independent of the effect of factor movements on domestic factor availability.

In the following section we view the *U.S.* economy as situated in the years 1963 and 1967 in an equilibrium such as *J*. A finding that the pattern of trade is consistent with preferences underlying (1a) or (1b) would lead to the conclusion that either previous capital-labor studies of the factor content of *U.S.* trade have encountered a persistent problem in their treatment of capital, or that our operating hypothesis of a composite energy-capital input is misplaced. We would obtain no "paradox" for capital, which would be observed to be consistently exported both directly and indirectly; but we would obtain a paradox with respect to energy, which would be observed to be imported directly and exported indirectly.

The alternative possibility is that the pattern of trade yields an outcome at *J* consistent with an income-expansion path (2). In that event energy would be consistently imported directly and indirectly, and hence would be unambigously indicated to be a scarce domestic factor. Further, the Leontief paradox would be substantiated, in that comparative advantage would be in the

[6] See Vanek (1968).

labor-intensive good, but capital, while imported indirectly, would be exported directly. However, our model allows the possibility in principle that labor services and capital services net of direct exports and commodity-embodied imports are both exported in exchange for the third factor, energy.

II. Derivation of Relative Factor Intensities of *U.S.* Tradeable Goods Production

To establish the relative factor intensities of *U.S.* tradeable goods production, the *U.S.* input-output tables for 1963 and 1967 at the 357 sector level of disaggregation[7] were converted to a domestic base by the subtraction from gross industry outputs of the c.i.f. values of competitive imports.[8] Discriminatory pricing manifested in declining block rate structures in industry energy sales causes dollar transactions data to be a poor indicator of sectoral energy employment. Hence for the five energy sectors (coal, crude oil and gas extraction, refined petroleum, and electric and natural gas utilities), the dollar transactions data were supplanted by physical energy input data.[9] On the supposition that energy and capital constitute a composite input, this substitution further provided a physical surrogate for capital inputs to be compared with the physical data for sectoral labor employment. The 1963 and 1967 tables with substituted energy sectors were then converted to matrices of technological input-output coefficients, $\Omega(1963)$ and $\Omega(1967)$. The resulting Leontief inverses $(I - \Omega_j)^{-1}$, j = 1963, 1967 were partitioned to form matrices of dimension (5 x 357) of direct plus indirect sectoral energy employment. The inputs for the five energy sectors were then aggregated in common British thermal units to establish a (1 x 357) vector of energy employment. Double counting is avoided by treating primary energy as the sum of coal, crude oil, and gas, and the fossil fuel equivalent of hydro- and nuclear-electric energy, so that for the ith sector the physical energy input obtained as

$$(15)\quad E_i = E_{1i} + E_{2i} + \beta E_{3i} - S_i \qquad i = 1 \dots 357$$

where β is a computational allowance for fossil fuel conversion efficiency and the fraction of nonfossil electric production,[10] and

$S_i = 1$ energy sector, $i = 1 \dots 5$

$S_i = 0$ nonenergy sector, $i = 6 \dots 357$

nets out for the energy sectors themselves their own intrinsic Btu content.[11] Premultiplication of the Leontief inverse by a vector of direct labor input coefficients[12] l yielded the corresponding sectoral direct plus indirect labor input vector

$$(16)\quad \lambda = \{l_i(I - \Omega)^{-1}\} \qquad i = 1 \dots 357$$

Equations (15) and (16) may be employed in two alternative approaches to the definition of the traded goods aggregates A and B

[7]The 357 sector model employed here results from minor restructuring of the standard U.S. Department of Commerce 368 sector tables. Special industries, scrap, Commodity Credit Corporation, and import sectors are deleted, while government enterprises whose output consists solely of secondary products are aggregated with the corresponding private sectors. For a detailed description of this model, see Bullard and Robert Herendeen.

[8]Competitive imports are defined as all "transferred imports" by the Bureau of Economic Analysis (*BEA*), plus competitive final goods directly allocated to final demand. (Source: *BEA* worksheets furnished by Robert Mangen of *BEA*.)

[9]In a simultaneous independent study, Norman Fieleke calculated the energy content of *U.S.* trade in 1970, but assumed energy transactions proportional to price in computing indirect requirements. Although the aggregate results thus obtained are quantitatively similar to ours, it has been shown by Herendeen that such an assumption can lead to errors up to 100 percent in energy intensities of certain goods. The effect on the aggregate results that one obtains therefore depends upon the structure of the arbitrary body of goods examined.

[10]This is consistent with commonly employed definitions of total energy consumption. The rationale is that hydro and nuclear electricity displace a certain amount of fossil fuel use.

[11]That is, we subtract the caloric content of the fuel itself (1 Btu/Btu) for the energy sectors.

[12]Direct labor data were assembled from U.S. Bureau of Labor Statistics publications. See Roger Bezdek et al.

identified in Section I. Exports and imports may be defined either inclusive or exclusive of intraindustry trade. Leontief's procedure, and that also followed by Baldwin, entails viewing import-competing and export goods as composites constructed with gross sectoral trade weights, and hence intraindustry trade is included in the calculation of relative factor intensities of traded goods. For given vectors of sectoral imports M and exports X, the Leontief mean factor intensities are computed as

$$(17)\qquad \bar{\alpha}_M = \frac{M'E_m}{M'\lambda_m},\quad \bar{\alpha}_X = \frac{X'E_x}{X'\lambda_x}$$

where for the jth sector we may have both $M_j > 0$ and $X_j > 0$ as respective elements in M and X, and the sectoral import and export energy-employment vectors E_m and E_x and labor-employment vectors λ_m and λ_x are respective partitions of the vectors given by (15) and (16). Computation of these means yielded

	$\bar{\alpha}_M$	$\bar{\alpha}_X$	$\bar{\alpha}_M/\bar{\alpha}_X$
1967	1451	1041	1.39
1963	1249	965	1.29

(values expressed in terms of millions of Btu/job)

The results indicate that in the decade of the 1960's *U.S.* international trade supplemented domestic energy availability via the factor content of traded goods. There was a general increase in the energy intensity of production between 1963 and 1967: the mean energy intensity of import-competing output increased 16 percent (from 1249 to 1451 million Btu/job) and the mean energy intensity of export production increased 7 percent (from 965 to 1041 million Btu/job). The differential in these changes raised the relative energy intensity of import competing over export production from 29 percent in 1963 to 39 percent in 1967.

The alternative course is to define import-competing and export aggregates net of intraindustry trade.[13] Each industry is then identified as a component of either import-competing or export output, and we obtain the trade-weighted relative energy-intensity distributions for net imports $\tilde{M}_i$,

$$(18)\qquad \xi_i \equiv \theta_i\tilde{M}_i = \frac{E_i}{\lambda_i}\tilde{M}_i \qquad \begin{array}{l} i = 1,\dots,118\ (1967)\\ i = 1,\dots,106\ (1963)\end{array}$$

and for net exports $\tilde{X}_i$,

$$(19)\qquad \xi_i \equiv \theta_i\tilde{X}_i = \frac{E_i}{\lambda_i}\tilde{X}_i \qquad \begin{array}{l} i = 1,\dots,209\ (1967)\\ i = 1,\dots,221\ (1963)\end{array}$$

The distributions (18) and (19) may be approximated by the continuous density functions $f_M(\xi)$ and $f_X(\xi)$, which yield the cumulative frequency distributions

$$(20)\qquad F_M(\xi) = \int_0^{\xi} f_M(T)dT$$

$$(21)\qquad F_X(\xi) = \int_0^{\xi} f_X(T)dT$$

When (20) and (21) are plotted against one another, the result for both 1963 and 1967 is that, with the exception of an initial low range of energy intensities, the export distribution stochastically dominates the import-competing distribution. That is, by stochastic dominance, import-competing production is relatively energy intensive. The means of the distributions (20) and (21) are

$$(22)\qquad \tilde{\alpha}_M = \frac{\sum_i \theta_i\tilde{M}_i}{\sum_i \tilde{M}_i},\quad \tilde{\alpha}_X = \frac{\sum_i \theta_i\tilde{X}_i}{\sum_i \tilde{X}_i}$$

which when evaluated yield

	$\tilde{\alpha}_M$	$\tilde{\alpha}_X$	$\tilde{\alpha}_M/\tilde{\alpha}_X$
1967	1667	1273	1.31
1963	1599	1223	1.31

(millions Btu/job)

That is, for both 1967 and 1963 the mean industry in the import-competing sample was 31 percent more energy intensive in

[13]For a study on this basis, see J. M. Finger.

production than its counterpart in the export sample.[14]

III. Natural Resources

Although Vanek's (1957) natural resource-capital complementarity argument does not in principle compromise the Heckscher-Ohlin Theorem (see William Travis and Vanek, 1968), nevertheless, elimination of natural resource sectors from Leontief and Baldwin's data samples significantly influenced their results; Leontief reversed the direction of relative factor intensity and Baldwin almost did the same. Baldwin further concluded that once one allows that factor-price equalization might not obtain, then the role of natural resources must be assigned prime significance in an explanation of the failure of the *U.S.* capital-labor data to support a factor proportions account of the determination of comparative advantage.[15]

In order to ascertain the possible effects of natural resource sectors on the results reported in Section II, relative factor intensities were reestimated with natural resources excluded. The initial consideration in this exercise is the specification of a consistent basis for industry selection. Robert Stern's warning is well taken, that "[b]ecause of the importance of natural resource products, it is difficult to know where to draw the line in defining industries for analytical purposes. Tests of the Heckscher-Ohlin model may therefore be confounded on both the export and import sides unless the investigator takes pains to recognize this issue" (p. 11). Our basis for sectoral distinction is cooperation with nature: all sectors wherein the indication was that fertility, climate, physical location of raw materials relative to the surface, or differential qualities of extracted materials in refining was prominent, were categorized as natural resource sectors. The intent was in this manner to eliminate from the industry samples those sectors which might most obviously not conform to the Heckscher-Ohlin assumption of internationally identical production functions. At the same time, it is industries in which factors are prominently engaged in cooperation with nature which might most be expected to depart from the assumption of identical sectoral energy intensities of capital equipment.

For the designated natural resource industries,[16] the direct inputs of labor and energy were set equal to zero, and trade of those sectors was eliminated from the import and export vectors, so that

$$L_j = A_{ij} = M_j = X_j = 0 \qquad (23)$$
$$j = \text{natural resource sector}$$
$$i = \text{energy}$$

This had the effect of deleting natural resource trade and of distributing the energy

[14]Actual plots of the distributions (20) and (21) are contained in the authors.

[15]See Baldwin (1971, p. 142).

[16]In the appendix to the authors' discussion paper, we provide a listing of the industries which we identified as natural resource sectors and a comparative cross-reference to the sectors eliminated on natural resource grounds by Leontief and Baldwin. Our criterion yields twenty-six natural resource sectors, compared with twenty-nine identified by Leontief and forty-two by Baldwin. The greater number of sectors chosen by Baldwin was partly the consequence of his use of a more aggregative model. Thus, for example, he found it necessary to include lumber and wood products together with logging. Baldwin also included, in addition to a set of natural resource products similar to those of Leontief, cotton, feed grains, and refined petroleum. Although we obtain approximately the same number of natural resource selections as Leontief, our selections are common only in sixteen cases: in the absence of specific criteria, Leontief's system of natural resources is hard strained for consistency; for example, crude oil and primary copper were natural resource products, but coal and primary aluminium were not. Vanek (1957) calculated the natural resource content of traded goods defining natural resource sectors as those using land for "extraction" rather than simply as a "site" for manufacturing. The natural resource criterion which we employ is broader than Vanek's, including not only his sectors, but also others where we would suppose energy-labor input proportions to be influenced by direct natural resource inputs. We include, for example, not only ores, but also primary metals, for the technology of ore extraction is affected by natural resource factors such as accessibility and concentration; but also the concentration of the extracted ore determines the amount of energy required to reduce it to metal.

and labor inputs of the domestic natural resource sectors free to purchasing sectors.[17]

The recalculated Leontief means (17) excluding and including natural resources are

		$\bar{\alpha}_M$	$\bar{\alpha}_X$	$\bar{\alpha}_M/\bar{\alpha}_X$
1967	*NR ex*	1255	1071	1.17
	NR in	1451	1041	1.39
1963	*NR ex*	1136	1047	1.09
	NR in	1249	965	1.29

(millions Btu/job)

Hence the exclusion of natural resource sectors from the industry samples consistently reduces the mean energy intensity of import-competing output and increases that of export production. This implies that the natural resource components of import-competing output were more energy intensive than the complete import-competing industry sample, and conversely, the natural resource components of export production were less energy intensive than export production as a whole. Eliminating natural resources from the data thereupon reduces the relative energy intensity of import-competing production, from 29 to 9 percent in 1963 and from 39 to 17 percent in 1967. By comparison, when Baldwin eliminated his natural resource designated sectors from the 1958 production 1962 trade data, the relative capital intensity of import-competing production declined from 27 to 4 percent.

When import-competing and export aggregates are identified on the alternative basis of net industry trade flows, the elimination of natural resources does not alter the stochastic dominance of the export distribution reported for the complete trade sample.[18] However, the effect of natural resource deletion is a diminution of the demarkation between the import-competing and export distributions. The recalculated distribution means excluding and including natural resources are

		$\tilde{\alpha}_M$	$\tilde{\alpha}_X$	$\tilde{\alpha}_M/\tilde{\alpha}_X$
1967	*NR ex*	1324	1173	1.13
	NR in	1667	1273	1.31
1963	*NR ex*	1461	1183	1.23
	NR in	1599	1223	1.31

(millions Btu/job)

Elimination of natural resources reduces the relative energy intensity of the mean import-competing industry, which is qualitatively consistent with the change in the Leontief mean; but now the energy intensity of the mean export industry is also reduced. As with the Leontief measure, however, the net consequence of deleting natural resources is once more a decline in the relative energy intensity of import-competing production, from 31 to 23 percent in 1963 and from 31 to 13 percent for 1967.

IV. Conclusion

Applying our model as a stylized description of *U.S.* foreign transactions yields the conclusion, after empirical investigation, that the United States exhibited a comparative advantage in labor-intensive output and a converse disadvantage in output intensive in a composite energy-capital input. The model further indicates such advantage in international trade does not have the consequence of a switch in the commodity direction of trade due to factor mobility. The results are very robust. The United States appears as having a comparative advantage in labor-intensive output when traded goods sectors are defined in the Leontief manner inclusive of intraindustry trade, and when sectors are identified on a directional trade basis by their net trade balances. Further, the results obtain consistently for both 1963

[17]It is quite possible that a lack of adequate computational facilities is the reason this more exact method was not employed in earlier papers. Besides recent improvements in computer hardware, software developments (see Killion Noh and Ahmed Sameh) vastly simplify the task of reinverting a 357 order matrix after such changes in Ω. Our computational procedure for eliminating natural resource sectors from the factor-intensity estimates is similar to that of Leontief, who treated natural resources as noncompetitive imports. It differs from Baldwin who apparently excluded natural resource products from trade flows but did not account for indirect input effects either by equalizing direct factor intensities or by treating natural resources as noncompetitive imports.

[18]For the distributions (20) and (21) with natural resources eliminated, see the authors.

and 1967. Deleting natural resource sectors diminishes the factor-intensity demarkation between domestic import-competing and export production, but for traded goods specifications both inclusive and exclusive of intraindustry trade and for both years investigated, the qualitative conclusion of *U.S.* comparative advantage in labor-intensive production remains confirmed.

An alternative finding that import-competing production was relatively labor intensive would have permitted us to question, through the medium of energy's role as surrogate measure of capital inputs, the validity of the capital-labor results underlying the Leontief paradox. Labor would have been imported indirectly, and capital would have been exported indirectly in addition to being observed to be exported directly. Our results, on the contrary, substantiate the findings of Leontief and Baldwin that import-competing production is relatively labor intensive. Our conceptual frame of reference with three inputs and two outputs offers, however, a more general perspective for the statement of conclusions. The presence of the third input, energy, permits an interpretation of the outcome of the perceived relation between factor endowments and intensities without recourse to the concept of paradox.

The United States engaged in international trade to conserve domestic endowments of energy, and energy was also imported as a direct factor flow. Hence energy is unequivocally identified as a relatively scarce domestic factor. Labor was exported via the factor content of traded goods and so is indicated as abundant relative to energy. For energy and labor the net factor flow directions are thus unambiguously determined; but we observe capital to have been imported indirectly and exported directly. In principle, the direct export of capital could have exceeded capital's commodity-embodied imported factor surplus. In the absence of credible capital data, and without a definitive means of evaluating international real capital transfers, both from the structure of the capital good content of foreign trade and from foreign investment recorded on the capital account of the balance of payments, we lack the necessary information for a direct test of this proposition.[19] However, we note that the Heckscher-Ohlin Theorem and our factor-intensity results are consistent in principle with the prefactor movement, original endowment rankings:

$$L^0/L^* > K^0/K^* > E^0/E^* \quad (24)$$

$$K^0/K^* > L^0/L^* > E^0/E^* \quad (25)$$

The ranking (24) embodies the relative factor endowment inference from the trade pattern known as the Leontief paradox, whereas (25) does not. Since (24) and (25) are both consistent with the data, we can only suggest that (25) is more reasonable in terms of informed expectations. This assigns to capital relative abundance, to energy relative scarcity, and to labor an intermediate status. It leads us to the inference that, with energy explicitly identified as a third factor of production, the net factor flows of capital and labor may have taken place in the same direction. The evidence indicates that in the 1960's, because of relative scarcity of energy, the net factor balance of *U.S.* international trade may have entailed the export of labor and capital services in exchange for imports, directly and indirectly, of energy.

Finally, our three-factor discussion has rested heavily on the presumption of technological complementarity between energy and capital, and the consequent inference that energy inputs may be employed as a surrogate for the factor services of capital.[20]

[19] On the problems entailed in the measurement of capital and the interpretation of capital data, see Vernon Smith, and on problems entailed in the identification of the nature of international capital transfers, see Baldwin (1970), and Andrew Schmitz and Peter Helmberger.

[20] Our focus on complementarity also suggests a relation to Vanek's (1959) proposed explanation of the Leontief paradox. Vanek suggested that complementarity between natural resources and capital, and domestic scarcity of the former and abundance of the latter, would lead more labor and less capital to be embodied in *U.S.* exports than a simple two-input capital-labor model would predict. As Travis and Vanek (1968) subsequently pointed out, if factor-price

If energy-capital complementarity were a technological fact and if there were no measurement errors in the evaluation of capital and energy inputs, then energy-labor intensities would correspond to capital-labor intensities; since no capital-labor intensity results are currently available for 1963 and 1967, we have no exact basis for comparison. Our energy-labor (Leontief) measures of 1.29 (1963) and 1.39 (1967) compare, however, with estimated capital-labor measures of 1.30 (1947, Leontief) and 1.27 (1957, Baldwin).

equalization obtains, the introduction of a third input complementary to one of the initial inputs does not reverse the comparative advantage predictions of the Heckscher-Ohlin Theorem. Hence the demonstration of input complementarity is not sufficient to explain the Leontief paradox.

REFERENCES

R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, *61*, 126–46.

———, "International Trade in Inputs and Outputs," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 430 34.

———, "The Role of Capital Goods in the Theory of International Trade," *Amer. Econ Rev.*, Sept. 1966, *56*, 841 48.

E. R. Berndt and D. O. Wood, "Technology, Prices and the Derived Demand for Energy," *Rev. Econ. Statist.*, Aug. 1975, *57*, 259–68.

Roger Bezdek et al. "Derivation of the 1963 and 1967 Total Employment Vector for 362 I–O Sectors," Document no. 63, Center Adv. Computation, Univ. Illinois-Urbana, Apr. 1973.

G. H. Borts, "A Theory of Long-Run International Capital Movements," *J. Polit. Econ.*, Aug. 1964, *72*, 341–59.

C. W. Bullard III and R. A. Herendeen, "Energy Impact of Consumption Decisions," *Proc. Inst. Electrical and Electronics Engineers*, Mar. 1975, *63*, 484–93.

N. S. Fieleke, "The Energy Content of U.S. Exports and Imports," disc. paper no. 51, Div. Int. Finance, Board of Governors, Federal Reserve System, May 1975.

J. M. Finger, "Factor Intensity and 'Leontief-type' Tests of Factor Proportions Theory," *Econ. Int.*, Aug. 1969, *22*, 405–22.

N. Georgescu-Roegen, "Energy and Economic Myths," *Southern Econ. J.*, Jan. 1975, *41*, 347–81.

R. A. Herendeen, private communication, Center Adv. Computation, Univ. Illinois-Urbana, Mar. 1975.

A. L. Hillman and C. W. Bullard III, "Energy, The Heckscher-Ohlin Theorem and U.S. International Trade," paper no. 25–77, Foerder Inst. Econ. Res., Tel-Aviv Univ. 1977.

R. W. Jones, "International Capital Movements and the Theory of Tariffs and Trade," *Quart. J. Econ.*, Feb. 1967, *81*, 1–38.

M. C. Kemp, "The Gains from International Trade and Investment: A Neo-Heckscher-Ohlin Approach," *Amer. Econ. Rev.*, Sept. 1966, *56*, 788–809.

W. W. Leontief, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, *38*, 386–407.

——— "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. Amer. Philosophical Soc.*, Sept. 1953, *97*, 332–49.

K. Noh and A. Sameh, "Computational Techniques for I–O Econometric Models," document no. 134, Center Adv. Computation, Univ. Illinois-Urbana, Oct. 1974.

J. Robinson, "Capital Theory up to Date," *Can. J. Econ.*, May 1970, *3*, 309–17.

A. Schmitz and P. Hemberger, "Factor Mobility and International Trade: The Case of Complementarity," *Amer. Econ. Rev.*, Sept. 1970, *60*, 761–67.

V. L. Smith, "The Measurement of Capital," in *Measuring the Nation's Wealth*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 29, New York 1964.

R. M. Stern, "Testing Trade Theories," in Peter B. Kenen, ed., *International Trade and Finance*, Cambridge 1975, 3–49.

William P. Travis, *The Theory of Trade and Protection*, Cambridge, Mass. 1964.

J. Vanek, "The Factor Proportions Theory: The N-Factor Case," *Kyklos*, Oct. 1968, *21*, 749–55.

———, "The Natural Resource Content of Foreign Trade, 1870–1955, and the Relative Abundance of Natural Resources in the United States," *Rev. Econ. Statist.*, May 1959, *41*, 146–53.

U.S. Department of Commerce, Bureau of Economic Analysis, *Input-Output Structure of the U.S. Economy*, Washington 1963; 1967.

Optimal Fiscal Reform of Metropolitan Schools: Some Simulation Results

By ROBERT P. INMAN*

In 1971 the California Supreme Court opened the door to a major reform movement to restructure the present system of decentralized school finance. With the exception of Hawaii, elementary and secondary education in the United States is supported primarily by local property taxation supplemented in part by state funded grants-in-aid. The California Supreme Court, in the now famous Serrano rulings, declared the California system in violation of the state constitution's equal protection clause. Similar rulings have also been handed down by the New Jersey Supreme Court (Robinson vs. Cahill) and the Superior Court of Hartford, Connecticut (Horton vs. Meskill). In addition, ten states have recently enacted major reform bills, and legislation is under consideration in several others. The pressure for reform is strong and continuing.

As a review of the recent reform proposals indicates, the legislative search for new means of financing local schools is not simply an incremental tinkering with existing laws.[1] Major changes, often court required, are at issue. Long-run outcomes are uncertain; each proposal has new winners and new losers. When planning a major reform of local school finance, therefore, past experience from incremental policymaking may not be an adequate guide to choice. Long-run general equilibrium predictive models and a clearly specified evaluation rule will be needed. It is the purpose of this paper to develop such a policy framework and to apply the analysis to one region currently in the midst of school reform, the New York metropolitan area. Six alternative reform proposals are considered: foundation aid, two district power equalization plans, property tax credits, expanded Title I assistance under the Elementary and Secondary Education Act, and centralized financing and spending controls. Preferred reforms are selected under utilitarian (pro-middle class), Rawlsian (pro-poor), and equal school spending (Serrano) criteria.

I. A General Equilibrium Model of the Metropolitan School Economy

Of crucial importance for evaluating alternative school reform measures is their impact on the levels and distribution of children's education and families' after-tax incomes. The key behavioral link which connects school fiscal policy to education and income is the demand for education. In this section I specify a model of educational demand which (i) details the behavioral impacts of the policy instruments for fiscal reform, (ii) explicitly incorporates long-run changes in local property tax base due to fiscally motivated family relocations, housing stock adjustments, and changes in firm investment, and (iii) allows for families to exit from the public school system for private education.

A. *Demand for Public Education*

To predict the general equilibrium effects of fiscal reform on local public education, descriptive models of school costs, taxable base, and the educational spending decision are required. The metropolitan economy with a center city and contiguous suburbs is selected as our "market boundary" as it is

*Associate professor of finance and economics, University of Pennsylvania. Financial support from the National Science Foundation (GS-44280) and the Spencer Foundation is gratefully acknowledged. This research has benefited from seminars at Duke, Harvard, and Penn. Doug Wolf's programming skills eased the research burden considerably. Julius Margolis first introduced me to the issues discussed here and has been a most helpful guide throughout.

[1]For a summary of recent revisions, see Education Commission of the States, *Major Changes in School Finance*.

within this economy that the major interactions following fiscal reform are likely to occur.

For this analysis I assume that the level of educational services per child (E) can be supplied at a constant cost per child and that all school districts are equally efficient in providing E. I denote the cost per child of providing E units of education as $k(E)$. Assuming further that the marginal and average costs per child of increasing E are equal over the relevant output ranges ($k(E)/E = k'(E)$), we can define educational output in expenditure units ($k(E)/E \equiv 1$) with no further loss of generality.[2]

To pay the costs of education, the school district is assumed to use only a tax against the market value of residential and commercial/industrial property. The *effective* local tax rate (r) is assumed to be proportional against all property values in each community, but of course the rate can vary from town to town.[3] The tax rate is set to insure that property tax revenues meet the per child costs of providing E minus any matching aid (at a percentage rate m) or lump sum aid (at z per child). If B is the per child market value of the local tax base then r is defined from the budget identity as

$$r \equiv \frac{k(E)(1 - m) - z}{B} \tag{1}$$

Given $k(E)$, B, m, and z, each community must select a level of E and, in turn, r. I shall assume that a consensus on E and r emerges which reflects the preferences of a decisive subset of community voters (for example, the median voter with majority rule) and that this decisive subset remains decisive through the economy's adjustment to any reform of school finances. The model in effect "freezes" the local political structure. While clearly open to question—particularly from advocates of a pure Downsian-style politics—this specification of local fiscal choice seems to accord reasonably well with the facts of school budgeting.[4]

The preferences of this decisive resident-representative for education and private income defines a demand schedule for E:

$$E = g(\tau, z, I) \tag{2}$$

where $\partial E/\partial \tau < 0$, $\partial E/\partial z > 0$, $\partial E/\partial I > 0$, and where τ is the marginal tax cost to the decisive resident of E, z is exogenous aid per pupil, and I is the decisive resident's *before*-local-school-tax income. The marginal tax costs of E are simply the change in taxes which follow a change in E. I assume all decisive voters deduct local taxes from federal and state income taxes at their marginal tax rates (whose weighted average is ω) and, when available, receive property

[2] For recent evidence on the constant cost per child assumption, see Terence Wales and Werner Hirsch. Less firmly based is the assumption that all school districts are equally efficient in providing E. This neglects two facts, both of which suggest center city school costs may exceed suburban costs. First, if there are significant peer group effects on learning, those communities with more favorable socioeconomic mixes will have lower per child costs in providing equal E's. Second, if some communities must pay higher factor prices for equal inputs—most likely teachers—they too will have higher costs per child. Unfortunately we lack tight evidence on the relevant parameters to predict exactly the size of any cost bias against center city children. But recent work suggests the bias may not be too large. Joseph Antos and Sherwin Rosen have estimated the wage premiums needed to attract a white male suburban teacher to the center city as at most $1,400/year. Given an average center city teacher/pupil ratio of .04, the premium is $56/student. For a common $1,000/student bundle of school inputs, the suburbs pay $1,000 and the center city pays $1,056 or a premium of 5 percent. The production function work of Donald Winkler, Richard Murnane, and Anthony Boardman et al. on peer group effects imply a possible additional 5 to 10 percent cost premium for equivalent service levels of center city children. Yet even if per child school costs are 20 percent larger in the center city, the qualitative conclusions of this research would be little changed. Quantitatively, the equal school spending plans and the pro-poor, pro-center city proposals would be made to look more pro-poor to compensate for higher costs, but the pro-middle class, pro-suburban policies would be little altered. See Section IV.

[3] Assessment rates are assumed to be uniform across all properties. The analysis below generalizes to nonuniform assessment rates provided the *relative* assessment rates remain invariant to aid changes and population movements induced by such changes.

[4] For evidence that an identifiable and stable decisive subset seems to control the school budget—both in big cities and in small towns—see the papers by Roscoe Martin, Arthur Vidich and Joseph Bensman, Warner Bloomsberg and Morris Sunshine, and Robert Lyke.

tax credits at rate $\lambda (0 \leq \lambda \leq 1)$ per dollar of local taxes against their income tax obligations. While the rate ω of deduction-based relief is given by the federal and state tax codes, the rate of the property tax credit λ is a "free" parameter which can be set by policymakers for school fiscal reform. The combined effects of deductions and credits gives an out-of-pocket cost of a dollar of local taxes of $(1 - \omega - \lambda) = \pi$. If b is the decisive resident's own tax base, then net taxes will be πrb, and the marginal cost schedule will be

$$\tau = d(\pi rb)/dE = \pi b(dr/dE) + \pi r(db/dE) \tag{3}$$

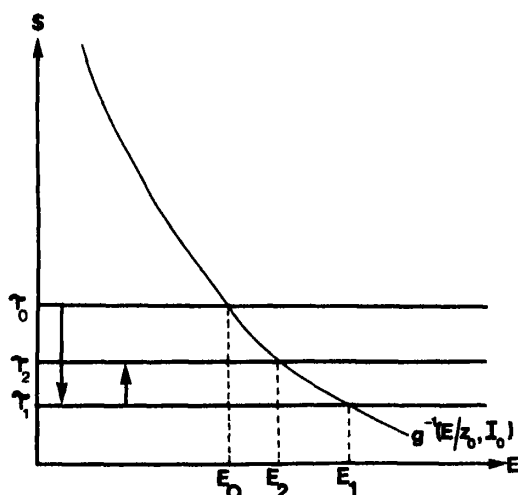
Empirically, the second term on the right-hand side of (3) is small and for clarity dropped here. (It is kept for the simulations, however.) Thus

$$\tau \simeq \pi(1 - m)b/B \tag{3'}$$

using (1) and our assumption (and normalization) that $k/E = k' \equiv 1$. Substituting (3') into (2) defines the locally preferred level of public education. The term E_0 in Figure 1 illustrates the equilibrium level of E for τ_0, z_0, and I_0. Given E_0 and (1) we can solve for r_0.

School fiscal reform directed at altering the levels and distribution of E and r operate in this model through one or more of the fiscal instruments: m, z, λ (and therefore π), and B. Each instrument has an obvious direct effect on E. Increasing m, B, or λ shifts τ downward leading to an increase in E to E_1. An increase in z shifts the demand schedule outward (not shown in Figure 1) which also increases E. The impact on the net tax rate πr is not so obvious, a priori. Increases in m, B, λ, and z all reduce πr, but the induced rise in E increases πr. It is an empirical matter, but recent evidence suggests aid increases are shared between increased expenditures and reduced taxes. These initial adjustments are the familiar story.

FIGURE 1

In fact, however, we should expect much more in a metropolitan economy of fiscally competitive school districts. Fiscal reforms which significantly alter the relative fiscal attractiveness of local communities will alter the pattern of family location, and, if housing sites are fixed in each community, local land values will change as well. Changes in local tax rates will alter family after-tax incomes and firm after-tax profits, leading to possible changes in housing and capital investments. These are clearly long-run changes, but they may have significant effects on local tax bases. Specifically b and B may change following aid reform leading to a long-run shift in τ, for example, from τ_1 to τ_2 in Figure 1 if the ratio (b/B) rises. Previous work on school fiscal reform and the grants literature generally has neglected these long-run adjustments in fiscal base.[5] This model incorporates them explicitly.

[5]In the education finance literature, see David Stern and Martin Feldstein. Strictly speaking Feldstein's interests were in identifying the structural facts needed for designing a wealth neutralizing aid formula rather than in predicting new school spending once his formula had been adopted. Had he pursued the prediction problem, he would have had to face the general equilibrium questions treated here. Eytan Sheshinski has specified a general equilibrium grants model which allows for housing stock adjustments, but it involves only one community so the capitalization feedbacks are necessarily omitted. George Meadows has built a capitalization model similar to the one presented here to test for the presence of Tiebout equilibria. I agree with his approach except that he uses changes in home and land values as synonymous, which is of course true only if there are no housing stock adjustments.

To do so, we need to specify the determinants of b and B.

The property base of the decisive resident in town i (b^i) is equal to the market value of his stock of housing plus the value of his land:[6]

$$(4) \qquad b^i = p_L^i + p_H \cdot H^i$$

where p_H is the cost of housing (housing stock is elastically supplied in all communities), p_L^i is the price of land in town i, and H^i is town i's decisive resident's stock of housing. Land values in town i are determined by the long-run availability of sites (S^i), local nonfiscal amenities (commuting distance, clean air) in towns i (Q^i) and towns j ($Q^{(j)}$, $j \neq i$), own fiscal position (E^i, r^i), and the fiscal position of neighboring communities defined by $B^{(j)}$, $m^{(j)}$, and $z^{(j)}$, ($j \neq i$):

$$(5) \quad p_L^i = \Phi(E^i, r^i; B^{(j)}, m^{(j)}, z^{(j)}; Q^i, Q^{(j)}; S^i)$$

I assume $\partial p_L^i/\partial E^i > 0$, $\partial p_L^i/\partial r^i < 0$, and $\partial p_L^i/\partial(\cdot)^j < 0$ ($\cdot = B, m, z$).[7] This is the now familiar capitalization equation implicit in recent studies of the Tiebout process.[8]

The decisive resident's demand for housing, measured as the stock H, is a function of the family's after-local-tax income (Y^i) and their effective price (construction plus tax costs) of housing (ρ^i):

$$(6) \qquad H^i = h(\rho^i, Y^i)$$

where $\partial H/\partial \rho < 0$ and $\partial H/\partial Y > 0$. The exact expressions for ρ^i and Y^i are somewhat involved and need not be reproduced here. (See the author, 1977a.) The main point to note here is that as the net tax rate πr^i increases, Y^i declines but ρ^i rises. Hence an increase (fall) in r^i leads to a fall (increase) in H^i.

The community's property tax base (B) is composed of residential property (Σb over all families in town i) plus commercial/industrial property. For this analysis, I assume firm relocation does not occur following school fiscal reform (I know of no convincing evidence one way or the other on this point), but that firms do adjust their investment decisions as local tax rates on capital are altered. Specifically, the commercial/industrial capital stock in town i (K^i) depends on r^i as:

$$(7) \quad K^i = l(r^i, \ldots,) \qquad \partial K/\partial r^i < 0$$

These seven equations for each town plus each community's fiscal base identity constitute a long-run model for the metropolitan public school economy. The model specifies equilibrium values of τ^i, E^i, r^i, p_L^i, H^i, K^i, and b^i for each of the i communities, given the levels of pre-local-tax family incomes and policy determined values m, z, and λ.[9] The value of B may be either exogenously determined by policy through various tax base equalization schemes (see Section II) or defined endogenously as equilibrium residential plus commercial/industrial property.

As an example of how the long-run fiscal base adjustment process might operate, imagine that town i in Figure 1 receives a relatively large increase in its matching rate, m^i. The tax price falls initially from τ_0 to τ_1 and E rises to E_1. In addition, r is also likely to decline initially. As m^i has now risen relative to other $m^{(j)}$'s ($j \neq i$), town i will now be fiscally more attractive, and potential new residents will bid up land prices. In addition, the fall in r will stimulate housing improvements or expansion and firm investment. Thus, both the decisive resident's base (b) and the commu-

[6]As tax rate r is an *effective* tax rate equal to the nominal rate times the assessment rate, the resident's tax base is the market value of his property. See fn. 3 above. I assume all decisive residents buy an equal size plot of land; zoning restrictions and the fact that my suburbs are all commuting suburbs of New York City lead to approximately uniform average plot sizes across towns.

[7]The specification in (5) is not perfectly general as I have neglected the effects of tax-financed changes in $E^{(j)}(j \neq i)$ on p_L^i. Empirically, however, this effect is small. See the author (1977a).

[8]See for example Wallace Oates (1969, 1973), and most recently Meadows.

[9]The assumption that pre-local-tax incomes are exogenous is the usual assumption in Tiebout-type models such as this one. For a justification in the context of this small region model, see the author (1977a, fn. 5). Note that I have also implicitly assumed the number of children in each community is exogenous. This seems reasonable for our sample cities with their fixed supply of residential sites.

nity's base per child (B) are increased. The two changes have offsetting effects on τ^i; increases in b^i increase τ^i while increases in B^i reduce τ^i. But in communities where fiscal capitalization occurs and where housing investment is more sensitive than firm investment to changes in r, b^i will rise proportionally more than B^i causing (b^i/B^i) and therefore τ^i to rise, say to τ_2. An identical adjustment with all the signs reversed occurs for cities receiving a relatively small increase or reduction in matching aid. In this example, the long-run effects partially *offset the initial impact effects of aid on E.*

Similar long-run stories can be told for all fiscal reform measures, and the reader is referred to my 1977a paper for an extended discussion. That study shows the long-run adjustments are significant; E_1 is often 5 to 20 percent larger (smaller) than E_2 for communities receiving differentially large increases (reductions) in fiscal aid.

B. *Demand for Private Education*

One of the major worries of those planning the fiscal reform of metropolitan schools has been the fear that a heavily redistributive reform will drive richer families from the public system and into private schools. A major exodus will probably undo any initial equalizing effects of aid on school spending, with perhaps the added side effect of encouraging further income segregation in housing patterns as the rich retreat to private school enclaves. It is important therefore that we be able to predict the likely effects of fiscal reform on the decision to use private schools.

The model that we develop here is a pure economic model; families prefer public or private schools simply by the criterion of maximizing family satisfaction as a function of after-tax and private school income and education. Religious or ethnic preferences are not at issue. Our analysis assumes that a family who drops out of the public system must continue to pay local school taxes; there are no private school enclaves which allow a family to totally escape the obligation of local taxes.

In Figure 2, $v(p, Y)$ is the family's de-

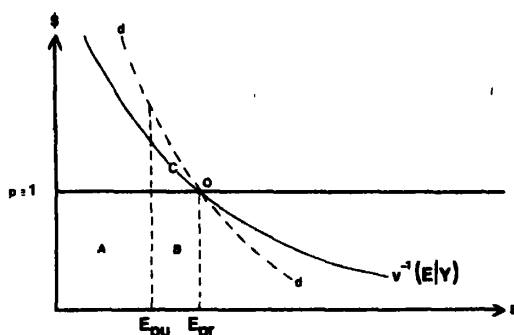



FIGURE 2

mand for education when purchased privately where p is the price of private schooling and Y is after-local-tax income. Here p equals the constant per unit costs of public schooling ($p \equiv 1$). For families in the private system with income Y the preferred level of private education is E_{pr} and disposable income after local taxes and private schooling is $Y_{pr} = Y - E_{pr}$.

How does the public school bundle (Y, E_{pu}) compare to the private school bundle (Y_{pr}, E_{pr})? As $Y > Y_{pr}$ always, the private school bundle can only be preferred if the family demands privately more education than their public system now provides ($E_{pr} > E_{pu}$). When the public school offers a family less than it would prefer to buy privately, the public system must "bribe" the family through increased private income ($Y > Y_{pr}$) if it is to remain within the public system. How big must the compensation be? The answer is the area under the compensated demand curve (dd) in Figure 2 through point 0 from E_{pr} to E_{pu}—area $C + B$. Does the school district offer the family compensation greater than $C + B$? Since the family must pay local taxes whether it uses the public school or not but private school tuition is paid only when the family uses the private school, the school district essentially offers families who choose E_{pu} rather than E_{pr} a savings of areas $A + B$, that is, the saved tuition cost of private schooling. Area $B + C$ is the compensation required before the family will accept E_{pu} while area $A + B$ is the school district's "offer." If $A + B > B + C$ (or $A > C$), then the public school system

is preferred. If $A + B < B + C$ (or $A < C$), then the family exits to the private system. For a given level of E_{pu}, area C is larger and exit is more likely as the family's demand curve for education ($v^{-1}(\cdot)$ in Figure 2) shifts outward. When education is a normal good, the rich are therefore more likely to leave the public system.

Using this model and given estimates of $v(p, Y)$, the fact that $p = 1$, and knowing E_{pu} and Y for each family in the region, we can estimate for each fiscal reform proposal both the number of public school families who exit to private schools and their chosen levels of E_{pr}. I make such calculations assuming each family takes as given its town's new long-run level of E_{pu} following reform, that there is no systematic shopping for private school enclaves, that private schools are available, and that public school districts do not adjust E_{pu} in response to wealthy families' threats of exit. The value of E_{pr} used as the basis of the consumer's surplus calculations is based on an estimated family demand for education schedule (see Section Ic below), an after-local-tax income level using the new, postreform spending and tax base levels in each district, and the assumption that price of private schooling equals 1. No dynamic feedbacks of exit back to spending are calculated. This exit model therefore gives only the *impact* effect of fiscal reform on private school use. If these impact estimates are large, say 30 percent or more of the region's children exiting to private schools, then the remaining public system may begin to unravel as private school enclaves form, as the loss of the rich leads to declining B's in nonenclave communities, as E_{pu}'s fall, and as the not-so-rich then find it advantageous to exit. If the initial losses are small, however, the impact estimates may be reasonably close to the true long-run decline in public school families.

C. *Empirical Specification*

The key equations of the public and private education models specified above were estimated for a portion of the New York metropolitan school economy—New York City plus fifty-eight Long Island school districts for the fiscal year 1968-69. The full details of the estimation process are described elsewhere (see the author, 1977a; 1978) but the central results and assumptions can be summarized here.

Family demands for elementary and secondary education (measured as instructional expenditures per pupil) were estimated using a variant of the median voter model proposed by Ted Bergstrom and Robert Goodman. The somewhat novel feature of the specification used here is that information on voter participation in school referenda is introduced, and I perform a direct test of the hypothesis that the family with the median income is the decisive voter. For my sample districts we cannot reject the null hypothesis that the median income family is decisive (see my 1978 paper). In each district then b^i, I^i, Y^i, and τ^i correspond to the family with the median income. The elasticity of E with respect to τ is estimated as $-.41$, the elasticity of E with respect to I as .60, and the elasticity of E with respect to z as .37. All parameter estimates are statistically significant at the .9 level and fall within the range of previous estimates.

An Oates capitalization equation was also estimated for our sample cities. The results are nearly identical to those in Oates (1973); full capitalization of exogenous shifts in own-local-school-tax rates (r^i) are indicated. From the Oates equation which uses median home value ($b^i = p_L^i + p_H H^i$) as the dependent variable we can infer land price changes (Δp_L^i) if we know the equilibrium effects of tax rate changes on housing stocks. These can be computed from a specification of the housing equation (see below). Cross-capitalization effects—the change in p_L^i as we change $B^{(j)}$, $m^{(j)}$, and $z^{(j)}$ for $j \neq i$—can also be computed from the Oates equation if we assume (i) a unitary price elasticity of demand for land (see Mahlon Straszheim, Table A.6), (ii) fixed and nearly equal average size of residential land sites, and (iii) that all towns are

equally competitive for residents.[10] The full specification of the capitalization model is described in my (1977a) paper.

The demand for residential housing follows Frank de Leeuw's estimates. The price elasticity of demand for H is -1.0 while the income elasticity of demand for the stock of housing is set at 1.4. The elasticity of commercial/industrial capital stock with respect to r^i is given a value of $-.25$, and I assume approximately 90 percent of taxable commercial base is held as capital (see Peter Mieszkowski).

Family demands for private education are based on the estimated median voters' demands for public education; schooling is schooling. The previously estimated median voter demand equation implies

$$E_{pr} = v(p, Y) = 5 \cdot p^{-.41} Y^{.6}$$

Area C is then measured by the triangular approximation $.5(E_{pr} - E_{pu})^2 \,|dp/dE|$, where $|dp/dE|$ is the absolute value of the slope of the compensated demand schedule and equals the inverse of the substitution term defined from $v(p, Y)$ at E_{pr}. Area A equals E_{pu}. Area C vs. area A comparisons were made for each family in the region assuming that the family income distribution in each town is stable and lognormally distributed.

The simulation algorithm used to approximate the new equilibrium values of public education and after-tax incomes (Y) is based on a first-order, linear approximation to the changes in public education and local tax rates from their existing levels. Private education and Y_{pr} follow from the demand for private schooling education given above. The details are provided in my paper with Wolf.

II. Aid Reform Proposals

Grants-in-aid available to our sample school districts for the fiscal year 1968–69 were dominated by state aid provided through a lump sum aid formula (called "foundation aid," see below) of the form,[11] $z \simeq 704 - .006B$. Districts also received a small amount of aid from federal matching aid programs (school lunch, vocational education, Title I); this aid was held fixed in the following analysis. As possible replacements for the existing New York school aid formula, we shall consider five commonly mentioned reform proposals plus variations in the state's present foundation aid plan.

Foundation Aid (FA): Foundation aid is the most prominent of current state-to-local school aid plans and awards each district a lump sum amount per child (\$$z$) defined by $z = z_0 - z_1 B$ ($z \geq 0$), where B is the market value of each district's per child tax base. The state sets a foundation level of school spending (z_0) and then requires each district to support that level at a tax rate of z_1. The district receives the difference between z_0 and their required contribution ($z_1 B$).

Foundation Aid with a Spending Limit (FALIM): The *FALIM* program considered here is that recommended by the President's Commission on School Finances. Each district receives a guaranteed (foundation level) amount of \$$z_0$ per child. The district is then free to supplement this amount from their own tax revenues up to 10 percent of z_0. Thus, $z_0 \leq E \leq z_0(1.1)$. Under this specification of *FALIM*, local school spend-

[10]The equally competitive assumption is probably not valid for New York City (*NYC*), but sufficient time-series data were not available to estimate a *NYC* capitalization equation directly. Our only recourse is sensitivity analysis. Two specifications were tried. (i) As a nonparticipant in the suburban land market. Fiscal changes are never able to move people into or out of *NYC*. Here $dp^i_L = 0$ for *NYC* and dp^i_L for the suburban communities is defined independent of *NYC* aid changes. (ii) *NYC* as viable but unique option for fiscally induced relocations with our capitalization effects of a *NYC* tax rate change equal to .5 the suburban own effects as estimated from the Oates equation. The results reported in Section III proved basically the same under the two specifications (see my 1977b paper), so only the simulations using specification (ii) are reported here.

[11]The foundation aid formula was estimated econometrically for our sample districts:

$$\underset{(18)}{z = 704} - \underset{(.0004)}{.006\,B} \quad R^2 = .71$$

(standard errors in parentheses)

ing is largely (at least 91 percent) state financed and is tightly controlled by the state through the choice of z_0. In effect, *FALIM* is a state financed school system.

District Power Equalizing Aid (DPE): The *DPE* scheme examined here guarantees to each community a target fiscal base per child (B^*) from which it can finance its school spending. The target base is protected by grants-in-aid which fill the gap between the revenues raised on own base and those raised at the same tax rate but with B^* as its base: Aid $= r(B^* - B)$. This scheme is a "linear" *DPE* plan in that it guarantees to each district a linear relationship between r and school resources per child. (Non-linear *DPE* schemes are proposed in John Coons et al.) The *DPE* linear aid formula is equivalent to the familiar percentage equalization formula, Aid $= (1 - B/B^*)$ (Tax Revenue), as $r =$ Tax Revenue$/B^*$. Note when $B > B^*$, Aid < 0; fiscally rich districts are required to contribute to the state aid pool.

DPE Aid with No Recapture (DPENC): This plan operates identically to *DPE* except that no districts receive a negative aid level— that is, there is no aid "recapture." For districts where $B > B^*$, Aid $= 0$. Most *DPE* formulaes now in use are of this form.

Matching Aid (MA): With *MA* the state agrees to share the costs of schooling at the rate of m per dollar of local spending, $0 \leq m \leq 1$. I adopt a Title I specification of matching aid by allowing m to vary inversely with district mean family income: $m = m_0 - m_1$ (District Mean Income).[12]

Tax Credit (TC): Of the six proposals, *TC* is the only plan which gives fiscal assistance directly to families. A property tax credit gives each family a credit against their state income tax of \$$\lambda$ per dollar of local school taxes. If local property taxes are also deductible from federal income taxes (at the marginal tax rate ω) then each family in effect pays \$$\pi(= 1 - \omega - \lambda)$ per dollar of local school taxes. For these simulations $0 \leq \lambda \leq (1 - \omega^u)$, where ω^u is the highest marginal tax rate of a tax deducting family. The upper bound insures that school tax relief does not become a de facto income subsidy ($\pi < 0$).

Financing Fiscal Reform: As each of the aid policies above involves a transfer of taxable resources from one community to another, the state government must have a taxing mechanism to affect these transfers. I assume that fiscal reform is paid for through a proportional state tax on income (I) and that this tax has no work disincentive effects. The financing of aid by taxing income raises additional feedbacks in the model. Changes in income (I) can change public school spending, housing stocks, and the relative attractiveness of private schools. These long-run effects of aid financing are explicitly incorporated into the analysis.

III. Selecting Preferred Reforms

The school fiscal reform debate is a prime example of a nonincrementalist policy debate. Recent state court rulings or the threat of judicial action has spurred several states (for example, California, New Jersey, Connecticut, and New York) to begin major redrafting of their state aid laws for financing local schools. Small changes of existing aid schemes are not legally or politically acceptable alternatives. With large policy changes, however, legislative intuition based on past experience may no longer be a sound guide to preferred policies. Past winners may now be losers and old losers can become winners. In such situations a formal evaluation rule may be needed to assist legislators through the maze of reform options. The rule will require a clear articulation of legislator preferences. No mean task to be sure, but if done with care and applied with caution the advantages for reasoned public choice are sizeable.[13]

[12] Feldstein's analysis of wealth neutrality proposals suggests relating m to each district's fiscal base. I do not consider such proposals explicitly here, but for my sample communities fiscal base and mean income are closely related implying our *MA* analysis based on income gives a reasonable approximation to effects of Feldstein's formula (suitably scaled) based on fiscal base.

[13] Abram Bergson sees this approach as the appropriate role for use of the social welfare function: "... We may usefully refer to it [*SWF*] if we are clear first

A. *Evaluation Rules*

For this analysis I assume that legislators' preferences over school reform depend on the levels and distribution across families in the region ($f = 1, \ldots, F$) of after-education family income (y_f) and per child education (measured as instructional expenditures per pupil). After-education family income is defined as before-reform income (I) *minus* local school taxes *minus* state aid reform taxes *minus* private school expenditures (= tuition) *plus* the annual value of the capitalized changes in the value of the family's residential plot.

The legislator's preferences are assumed to be reducible to the evaluation rule:

$$(8) \quad W = \left\{ \sum_f U(y_f, E_f)^\gamma \right\}^{1/\gamma}, \quad \gamma \leq 1$$

where $U(y_f, E_f)$ is the planner/politician's evaluation of family f's postreform position and is specified here by $U = y_f^\Delta E_f^{1-\Delta}$ with $0 \leq \Delta \leq 1$. The parameter Δ defines the relative importance of education vis-à-vis after-education spendable income. The larger the legislator's value of Δ, the less important he or she views schooling as a determinant of family welfare. The parameter γ defines the legislator's relative aversion to inequality in the distribution of the $U(y_f, E_f)$ When $\gamma = 1$, all families are weighted equally and the legislator's W reduces to the familiar Benthamite evaluation rule measuring regional welfare as the simple sum of family welfare. When γ is large and negative, W weights more highly those families with low values of $U(y_f, E_f)$. As $\gamma \rightarrow -\infty$, W collapses to the Rawlsian maximin criterion: select those policies which maximize the welfare of the worst off.[14] Given a legislator's values for Δ ($0 \leq \Delta \leq 1$) and γ (≤ 1) and our general equilibrium model of educational finances, we can ordinally rank all fiscal reforms by the rule W.

To illustrate the usefulness of this approach, I have ranked each of the six reform proposals of Section II against three rather different specifications of W. The first evaluation rule is a Benthamite utilitarian criterion (WU) setting $\Delta = .83$ and $\gamma = 1$. Thus WU favors the lower-middle and middle class families who account for most of the region's population.[15] A value of $\Delta = .83$ implies that the legislator would like to see, on average, each family devote about 17 percent (= $1 - \Delta$) of its prereform income to (state plus or private) educational spending, leaving 83 percent for other goods and services (y_f).[16] In 1970, the average New York metropolitan area family did spend approximately 17 percent of its income on state plus local school spending.[17] Thus, setting $\Delta = .83$ implies legislator approval of the average family's 1970 performance. The rule WU is probably a good indicator of a pro-middle class legislator's preferences.

The second evaluation rule is an approximate Rawlsian rule (WR) setting $\Delta = .83$ again but giving γ a value of -10, the

that welfare economics is envisaged essentially as a form of counselling.... The counsel consists of implications of some criterion of social welfare, that is, some values delimiting the 'social welfare' attaching to different social states, and hence providing a basis for ordering such states.... The counsel ... might be proffered to citizens generally, but is usually intended especially for public officials" (p. 186). On various approaches for inferring legislators' preferences, see Leif Johansen. For an approach consistent with the evaluation rules proposed here see the author and Wolf.

[14]See Sidney Alexander.

[15]In 1969 for the New York Standard Metropolitan Statistical Areas (*SMSA*), approximately 55 percent of the families had annual incomes between $5,000 and $15,000. As I am using approximately one-third of the total New York suburban school districts which surround the city, I scaled all New York City enrollment and family population numbers to one-third their true values to retain the true regional income distribution numbers.

[16]Maximizing $U = y_f^\Delta E_f^{1-\Delta}$ subject to constraint $I = y_f + E_f \cdot 2$ (each family has $\sim$ 2 school-age children) yields as the *legislator's* preferred family demand, $2 \cdot E_f = (1 - \Delta)I$.

[17]In 1970, mean family income in the New York metropolitan region was about $15,700 with mean state plus local spending for educational instruction of $1,165 per child. With an effective state plus federal tax rate for noneducational spending of $\sim .13$, our before-local tax I equals (15,700*.87) = $13,700. Thus 17 percent of I(= 1,165*2/$13,700) was devoted to instructional expenditures by the average family.

largest negative number the computer could employ when calculating WR. The function WR seems a reasonable approximation to a pro-poor legislator's preferences.

The final evaluation rule is a normalized variant of W designed to rank policies according to their equalizing impacts on the region's distribution of education. The normalization is achieved by dividing W in (8) evaluated at $\Delta = 0$ (only education matters!) by an "adjusted" mean level of school spending. The result is an Atkinson inequity measure for the distribution of education (see A.B. Atkinson) which I denote by WEQ:

$$WEQ = \hat{E}/\bar{E} \tag{9}$$

where $\bar{E}$ is the average level of E received by all children from public and private schools and $\hat{E}$ is the amount of E, which if it were equally distributed, would give an outcome which the decision maker would consider to be equally as good as the existing, possibly unequal distribution of E.[18] If the legislator shows any aversion to inequalities ($\gamma < 1$) and if the present distribution is anything but perfectly equal then $\bar{E} > \hat{E}$ and $WEQ < 1$. Generally, $0 \leq WEQ \leq 1$. For a given value of γ, the larger is WEQ the more equal is the observed distribution of education.[19] Policies which maximize WEQ should presumably appeal to supporters of Serrano.[20] For these simulations a moderately proequalizing value of $-.5$ was chosen for γ.

B. *Some Simulation Results*

Tables 1-3 summarize the preferred policies under WU, WR, and WEQ. First, the best policy parameter values were chosen for each of the six proposals listed in Section II.[21] The W maximizing parameter

[18] Assuming all F families have two children and letting $\bar{E} = \sum_f E_f/F$ and defining $\hat{E}$ by the equality

$$\{\hat{E}^\gamma\}^{1/\gamma} = W_{\Delta=0} = \{\Sigma E_f^\gamma\}^{1/\gamma}$$

we can easily show the normalization

$$WEQ = W_{\Delta=0}/(\Sigma E_f/F^{\gamma-1/\gamma})$$

[19] If decision makers dislike inequalities in the distribution of E, then presumably they would pay a premium in reduced $\bar{E}$ to see the distribution made more equal. Interpreting $\hat{E}$ as the "equality equivalent" of the distribution of E in a manner fully analogous to the usual interpretation of a distribution's certainty equivalent, then the size of this premium is measured by $\bar{E} - \hat{E}$. The index WEQ is just one minus the percentage reduction in E which the planner will forego to insure perfect equity: $WEQ = 1 - (\bar{E} - \hat{E})/\bar{E}$. The larger is WEQ, the smaller is the relative size of the decision maker's premium, and hence the "more equal" he or she views the distribution.

[20] There is some question whether Serrano and its progeny require equal spending or simply wealth neutrality. The wealth neutrality objective (school spending independent of local tax base) is analyzed nicely in Feldstein. I read the spirit of these cases as equal spending cases, however, with wealth neutrality the current vehicle to that end. If wealth neutrality does not give equality (and the results here suggests it may not), then the lawyers and any sympathetic courts will be back looking for new remedies. Table 3 below suggests new candidates.

[21] Policies were searched in the following intervals for the optimal parameters: B in \$1,000 units; z_0 in \$10 units and z_1 in .005 units; m_0 and m_1 in .01 units; and λ in .01 units.

TABLE 1—*WU* RANKINGS

Policy	Rank	Optimal Parameter Values	State Tax Rate	$\bar{E}$	*NYC/SUB*	Percent Exit
DPE	1 (Best)	$B^* = 37000$	0	930	.75	13.7
FA	2	$z_0 = 2150$	.039	1036	.57	10.3
		$z_1 = .05$				
TC	3	$\lambda = .30$	.018	913	.90	12.7
Existing Aid	4	$z_0 = 704$	.060	1165	.92	0
		$z_1 = .006$				
DPENC	5	$B^* = 44000$	.020	962	.72	11.3
MA	6	$m_0 = 0$	0	869	.78	22.0
		$m_1 = 0$				
FALIM	7 (Worst)	$z_0 = 1080$	.138	1187	~1	~0

TABLE 2—*WR* RANKINGS

Policy	Rank	Optimal Parameter Values	State Tax Rate	$\bar{E}$	*NYC/SUB*	Percent Exit
FA	1 (Best)	$z_0 = 40$ $z_1 = -.01$	.054	1152	.92	2.2
TC	2	$\lambda = .30$	.018	913	.90	12.7
Existing Aid	3	$z_0 = 704$ $z_1 = .006$	.060	1165	.92	0
MA	4	$m_0 = .17$ $m_1 = .02$	~0	872	.91	20.3
DPE	5	$B^* = 37000$	~0	930	.75	13.7
DPENC	6	$B^* = 13000$	0	869	.78	22.0
FALIM	7 (Worst)	$z_0 = 1060$	.132	1166	~1	~0

values along with required state aid tax rates are listed in Tables 1-3. (The supplemental state tax rates required to support each aid scheme may seem a bit high, but it should be remembered the aid plans are supported by only a proportional tax on residential income; state sales taxes or corporate taxes are not used here.) Also reported for each preferred policy are the resulting mean levels of spending on instruction ($\bar{E}$), the New York City to mean suburban school spending ratio (*NYC/SUB*), and the percent of children who leave the public system for private schools following reform (Percent Exit). Each of the six preferred proposals are then ranked by the computed values of *WU*, *WR*, and *WEQ*. As a point of reference the existing New York state aid plan is also included in Tables 1-3 and ranked against the best reform proposals.

Space limits detailed comments on these results (available in a longer version of this paper), but some highlights should be noted.

1) Surprisingly the district power equalizing proposals (*DPE* and *DPENC*) which have been so strongly pushed by lawyers and equity proponents (for example, Coons et al.) do very poorly under the equity (*WEQ*) and pro-poor (*WR*) criteria, but under the efficiency oriented utilitarian criterion (*WU*), *DPE* rises to the top of the list. Why?

The top performance of *DPE* against *WU* is attributable to its unique self-financing provision which increases the effective local school property tax rates in fiscally rich towns ($B > B^*$, Aid < 0) to finance school tax rate reductions in fiscally poor towns. When B^* is near the region's mean fiscal base per child, *DPE* becomes largely self-financing; no state tax on personal income is needed. This is exactly what happens when $B^* = \$37,000$/child (see Table 1). The net effect is sizeable tax relief for the many

TABLE 3—*WEQ* RANKINGS

Policy	Rank (*WEQ*)	Optimal Parameter Values	State Tax Rate	$\bar{E}$	*NYC/SUB*	Percent Exit
FALIM	1 (.9999)	$z_0 = 1080$	.138	1187	1.00	.5
MA	2 (.9939)	$m_0 = 3.55$ $m_1 = .16$	.102	968	.96	7.2
FA	3 (.9909)	$z_0 = 760$ $z_1 = .01$	.048	1074	.85	2.2
Existing Aid	4 (.9884)	$z_0 = 704$ $z_1 = .006$	.060	1165	.92	0
TC	5 (.9783)	$\lambda = .3$	.018	913	.90	12.7
DPENC	6 (.9737)	$B^* = 56000$	.044	1035	.72	6.1
DPE	7 (.9735)	$B^* = 56000$	.043	1038	.72	5.9

residents of the fiscally poor bedroom suburbs. And as their communities now face a lower equilibrium tax price for education, E rises too. The families who are hurt the most by *DPE* live in the fiscally rich residential suburbs. School tax prices rise significantly leading to a fall in both y and E. New York City families see a slight rise in tax prices as the city's initial tax base exceeds B^* (= $37,000) and educational spending falls from $1,123/child with current aid to $760/child with *DPE*. But the zero state tax rate for *DPE* provides New York City residents with significant tax relief over the present aid plan (using a 6 percent tax rate), and on balance the increase in y compensates for the fall in E. With *DPE* at B^* = $37,000/child, the large middle class gains, the rich suburbs lose, and New York City residents about "break even." Note that *DPENC* which constrains Aid ≥ 0 does not have the self-financing provision used by *DPE*. Hence its poor performance under *WU*.

For the same reason that *DPE* helps the suburban middle class, we also see that under *WR* it hurts, or does not help very much, the center city poor. Under *WR* the best *DPE* policy is to keep B^* at $37,000. The *DPENC*, which cannot give the poor in New York City any tax relief at all without drastic reductions in E, is a very poor *WR* performer; the reason is New York City's relatively high initial fiscal base of $43,300/child.

But why does *DPE* do so badly under *WEQ*, presumably the objective for which it was designed? The reason is apparent from Figure 1. Low-base, low-spending school districts originally see a fall in their tax prices with *DPE* (to say τ_1) and thereby increase E. High-base, high-spending districts see a rise in τ thereby reducing E (a short-run increase in equity). In the long run, strong capitalization and housing stock adjustments (increasing b in low-base towns, reducing b in high-base towns) return tax prices to near their starting values ($\tau_2 \rightarrow \tau_0$). The long-run result is a distribution of E not very different than the no-aid status quo. There are other non-*DPE* plans which do much better against *WEQ*.

2) Foundation aid appears to be a reasonably flexible reform policy, doing well in the ordinal rankings against all three criteria. But note the optimal parameter values change radically as we move from *WU* to *WR* to *WEQ*. Under *WU*, emphasis is on helping the middle class so aid goes only to those communities with equilibrium tax bases less than $43,000/child (= $2,150/.05). New York City does not get aid. Under *WR*, however, it gets first attention and the optimal aid formula is *FA* with a profiscal base bias—Aid = 40 + .01B. To help the poor, money must go to New York City, and as the city has a large fiscal base dominated by relatively insensitive commercial property, a pro-base policy emerges. Here the rich suburbs ride on the poor's coattails, while the middle class residential suburbs are the relative losers. In this school economy, when we help the poor we hurt the middle class. There may be ways to ease this tension—for example, foundation aid plans which are redistributive toward towns with both low incomes and low fiscal bases ($Z = z_0 - z_1 B - z_2$ Mean Income)—but no such proposals have been seriously considered as a basis for reform.

3) Matching aid is a poor performer under *WU* and *WR* but is a second best *WEQ* policy. The *MA* suffers from the strong offsetting capitalization and investment effects in this economy (see Figure 1). As a result it proves so expensive in state taxes to increase E by matching aid that the best strategy under *WU* and *WR* is to do nothing (*WU*) or almost nothing (*WR*, some small aid to very poor towns) at all! When state tax rates and y do not matter, as under *WEQ*, then *MA* has a role to play. Here aid has a strong equalizing flavor. Matching rates are zero for all towns with mean family incomes greater than $22,200 (= 3.55/.16 × 1,000) and 1 (full state funding) for all towns with mean family incomes less than $15,900. It is an expensive program.[22]

[22]Martin Feldstein has suggested to me, and I agree, that there may be a bias against *DPE* and *MA* policies due to the way I specified and estimated my demand for public education equation. My tax price τ was defined as b/B so that the major source of variation in district

4) In effect, foundation aid with a 10 percent spending limit is a state funding proposal for local education financed only by a proportional tax on residential income. As such we lose our option to tax commercial property. It is not surprising then that *FALIM* is the worst performer against *WU* and *WR*. Nor is it surprising that *FALIM* is the best education equity performer given its tight control over local spending. The only issue for *FALIM* under *WEQ* is to set the level of state support high enough to keep rich families in the public system and therefore under state control. At z_0 = \$1,080/child, exit is minimized and all but a few districts choose to spend at the allowable upper limit of \$1,188/child.

5) The tax credit operates as a subtle income redistribution device and as such it does well in the *WU* and *WR* rankings and poorly under *WEQ*. The net income transfer for a typical resident is his income tax credit (λrb) minus the increase in his state taxes needed to finance the plan ($s \cdot I$, where s is the state tax rate in Tables 1–3). As a household's taxable base is the value of land plus dwelling and empirically $b \sim 2.5\ I$ (see Henry Aaron, for example), we have a net tax savings equal to $(2.5\ \lambda r - s)I$. The family gains, breaks even, or loses as $2.5\ \lambda r \gtreqless s$. For the *WU* and *WR* optimal *TC* plans, families living in towns with local tax rates (r) greater than .024 are net gainers. These will generally be the poor and middle class families in New York City and in the fiscally poor suburbs.

values of τ comes from variations in commercial base. But if the commercial tax base is thought to be highly sensitive to increases in local school tax rates, towns may feel constrained in increasing E for fear of losing base. This sensitivity to the loss of commercial base would not be observed for truly exogenous changes in matching rates. Thus, we would expect my estimate of the tax price elasticity to be less in absolute value than estimates of the price elasticity derived from variations in matching rates. The work of Feldstein and Helen Ladd confirm this; they find matching aid elasticities nearer −.7. A higher absolute price elasticity will of course make *MA* and *DPE* more effective instruments for fiscal reform. Unfortunately, New York does not now have a matching aid program which would allow us to repeat the Feldstein-Ladd approach.

It should be mentioned that under all three objectives, λ wants to rise above .3, but .3 was the upper limit set by the constraint that $\pi = 1 - \omega^u - \lambda \geq 0$. In a sense the uniform tax credit (λ) seeks to undo the unequalizing effects of the regressive tax deduction (ω, which rises with income). Dropping tax deductions and introducing proportional or progressive tax credits may be a promising new policy option for all three objectives.

6) A comparison of the preferred policies from Tables 1–3 makes clear that there is no dominant proposal. Schemes which do well under one criterion do poorly under another. The implications of this conflict are important. First, the legislative debate over reform policies may quickly collapse to a pro-middle class vs. pro-poor debate. New, compromise proposals which can satisfy both proponents of the poor (legislators with *WR* preferences) and proponents of the middle class (legislators with *WU* preferences) should be found. Revised *FA* plans or progressive tax credits are two alternatives. Second, the conflict of *WU* and *WR* with *WEQ*—notably over the efficacy of *FALIM* and *MA*—presage a familiar story. Effective legislative responses to meet court-imposed standards of equity may be hard coming. High proportional state tax rates to finance equalizing fiscal reform (*FALIM* and *MA*) do not sit well (or long) with utilitarian or Rawlsian legislators. The two-year struggle in New Jersey over a school-aid package involving a new state income tax is a first case in point.

IV. Conclusion

The quantitative results of this paper should be used with caution for they are only one set of simulations for one regional economy. While the ordinal rankings of the policies and our qualitative discussion of the outcomes generally hold for other values of Δ (= .95 and .67) and model specifications, the optimal values of the individual policy parameters do change significantly.

The major shifts in the ordinal ranking of policies occur under *WU* for Δ = .67.

Here a *FA* scheme ($z_0 = 1920$; $z_1 = .01$) replaces *DPE* ($B^* = \$192{,}000$) as the preferred policy. The relative advantage of *FA* over *DPE* aid as an instrument for stimulating educational spending is the reason for *FA*'s dominance with the now larger preference weight on education. In addition, when $\Delta = .67$ the *DPENC* and *FALIM* policies move up the list of preferred utilitarian proposals. For $\Delta = .67$, the optimal *DPE* target base is $192,000 and all districts receive aid. Thus the recapture provision of *DPENC* is no longer operative and *DPE* and *DPENC* are effectively the same policies, both ranking second behind *FA*. When $\Delta = .67$, the large losses of private income which plagued *FALIM* at $\Delta = .83$ are no longer a serious limitation; the premium here is on increasing *E*. The *FALIM* can increase *E* by setting a high z_0. With $z_0 = \$1920$, *FALIM* moves to third behind *FA* and *DPE*. The *FA* remains the best policy under *WR*, both for $\Delta = .95$ ($z_0 = 2300$; $z_1 = .12$) and $\Delta = .67$ ($z_0 = 200$; $z_1 = -.04$).[23] Again at $\Delta = .67$ we see an elevation in the rankings of *FALIM* (now the second best policy at $z_0 = \$1{,}880$) and *DPENC* ($\sim$ *DPE*, with $B^* = \$167{,}000$). Changes in γ to .5 or to -1.0 for *WEQ* leave the optimal policy values, and therefore the *WEQ* rankings, largely unaffected. On balance the qualitative conclusions 1-6 remain valid even against a rather wide range of preference specifications.

Variations in the specification of the regional economy should also alter the preferred policy values. Some preliminary work in this direction indicates preferred parameter values change when we drop the private school model and assume no exit (see a longer version of this paper) and when we drop the long-run capitalization model (see my 1977a paper), but again the ordinal rankings and qualitative results are not much altered. The major result from the no capitalization model is the improved relative performance of *DPE* under *WEQ*, but as before *FALIM* and *MA* still dominate.

Thus while the qualitative conclusions seem robust across alternative specifications of the model and the politician's preference function, the final quantitative specifications of an optimal aid package are quite sensitive to these key parameters. There is no escaping the fact that sound fiscal planning requires sound empirical analysis. What works in New York may not work in Sheboygan.

[23]The striking switch in the best *FA* plan from a strongly redistributive fiscal proposal when $\Delta = .95$ ($z_1 = .12$) to strongly antiredistributive plan when $\Delta = .67$ ($z_1 = -.04$) is due to the shift in emphasis from private income to education. When $\Delta = .95$ the best *FA* plan under *WR* gives a small amount of aid to the few towns whose equilibrium values of *B* are less than $17,500. The state tax rate needed to support this scheme is nearly zero. The bulk of the poor families in the region therefore receive an increase in private income with the few poor families in the very fiscally poor towns receiving some education support. When $\Delta = .67$ the emphasis shifts to increasing the education levels of the poor. As the bulk of the poor live in industrialized, high base towns where we need to *increase* aid to increase *E*, a value of z_1 equal to $-.04$ is the answer.

REFERENCES

Henry Aaron, *Shelters and Subsidies*, Washington 1972.

S. Alexander, "Social Evaluation Through Notional Choice," *Quart. J. Econ.*, Nov. 1974, *88*, 597-624.

J. Antos and S. Rosen, "Discrimination in the Market for Public School Teachers," *J. Econometrics*, May 1975, *3*, 123-50.

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244-63.

A. Bergson, "Social Choice and Welfare Economics Under Representative Government," *J. Publ. Econ.*, Oct. 1976, *6*, 171-90.

T. C. Bergstrom and R. P. Goodman, "Private Demand for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280-96.

Warner Bloomberg and Morris Sunshine, *Suburban Power Structures and Public Education*, Syracuse 1963.

A. Boardman, O. Davis, and A. Lloyd, "A Si-

multaneous Equation Model of the Educational Process Reconsidered," presented at the Econometric Society Meetings, Winter 1974.

B. Brown and D. Saks, "The Production and Distribution of Cognitive Skills Within Schools," *J. Polit. Econ.*, June 1975, *83*, 571–93.

John Coons et al., *Private Wealth and Public Education*, Cambridge 1970.

F. de Leeuw, "The Demand for Housing: A Review of the Cross-Section Evidence," *Rev. Econ. Statist.*, Feb. 1971, *53*, 1–10.

M. S. Feldstein, "Wealth Neutrality and Local Choice in Public Education," *Amer. Econ. Rev.*, Mar. 1975, *65*, 75–89.

W. Hirsch, "The Supply of Urban Public Services," in Harvey Perloff and Lowdon Wingo, eds., *Issues in Urban Economics*, Baltimore 1968, 477–525.

R. Inman, (1977a) "Micro-fiscal Planning in the Regional Economy: A General Equilibrium Approach," *J. Publ. Econ.*, Apr. 1977, *7*, 237–60.

——— (1977b), "Optimal Fiscal Reform of Metropolitan Schools: Some Simulation Results," mimeo, Univ. Pennsylvania 1977.

———, "Testing Political Economy's 'As If' Proposition: Is the Median Income Voter Really Decisive?," *Publ. Choice*, forthcoming 1978.

——— **and D. Wolf,** "SOFA: A Simulation Program for Predicting and Evaluating The Policy Effects of Grants-in-Aid," *Socio-Econ. Planning Sci.*, June 1976, *10*, 77–88.

L. Johansen, "Establishing Preference Functions for Macroeconomic Decision Models," *European Econ. Rev.*, June 1974, *5*, 41–66.

H. Ladd, "Local Education Expenditures, Fiscal Capacity, and Composition of the Property Tax Base," *Nat. Tax J.*, June 1975, *28*, 145–58.

R. Lyke, "Representation and Urban School Boards," in Henry Levin, ed., *Community Control of Schools*, Washington 1970.

Roscoe Martin, *Government and the Suburban School*, Syracuse 1962.

G. Meadows, "Taxes, Spending, and Property Values: A Comment and Further Results," *J. Polit. Econ.*, Aug. 1976, *84*, 869–80.

P. Mieszkowski, "The Property Tax: An Excise or a Profits Tax?," *J. Publ. Econ.*, Apr. 1972, *1*, 73–96.

Richard Murnane, *The Impact of School Resources on the Learning of Inner City Children*, Cambridge 1975.

W. Oates, "The Effects of Property Taxes and Local Public Spending on Property Values," *J. Polit. Econ.*, Nov. 1969, *77*, 957–71.

———, "The Effects of Property Taxes and Local Public Spending on Property Values: A Reply and Yet Further Results," *J. Polit. Econ.*, July 1973, *81*, 1004–08.

E. Sheshinski, "The Supply of Communal Goods and Revenue-Sharing," in Martin Feldstein and Robert Inman, eds., *The Economics of Public Services*, New York 1977.

D. Stern, "Effects of Alternative State Aid Formulas on the Distribution of Public School Expenditures in Massachusetts," *Rev. Econ. Statist.*, Feb. 1973, *55*, 91–97.

Mahlon Straszheim, *An Econometric Analysis of the Urban Housing Market*, New York 1975.

A. Vidich and J. Bensman, "The Clash of Class Interest," in Alan Rosenthal, ed., *Governing Education: A Reader*, Garden City 1969.

T. Wales, "The Effect of School and District Size on Education Costs in British Columbia," *Int. Econ. Rev.*, Oct. 1973, *14*, 710–20.

D. Winkler, "Educational Achievement and School Peer Group Composition," *J. Human Resources*, Spring 1975, *10*, 189–204.

Education Commission of the States, research brief, *Major Changes in School Finance*, Vol. 2, No. 2, May 1974.

New York State Commission on the Quality, Cost, and Financing of Elementary and Secondary Education, *The Fleischmann Report on the Quality, Cost and Financing of Elementary and Secondary Education in New York State, Vol. I, New York 1973.*

President's Commission on School Finance,

Schools, People and Money, The Need for Educational Reform, final rept., *Washington, 1972.*

Horton vs. Meskill, 31 Conn. Superior Court No. 377, 332 A.2d 113, 1974.

Robinson vs. Cahill, 118 N.J. Superior Court No. 223, 1972; aff'd 62 N.J. 473, 303 A.2d 273, 1972; cert. denied 414 U.S. 976, 1973.

Serrano vs. Priest, L.A. 29820, Superior Court No. 93854; cited in *Harv. Educ. Rev.*, Nov. 1971, *41*, *503*.

An Analysis of the Changing Location of Iron and Steel Production in the Twentieth Century

By JOHN S. HEKMAN*

Historical changes in the pattern of industrial location are frequently associated with geographical differences in the cost of production. In the case of the iron and steel industry, several writers found apparent regional cost differences which suggested reasons for the westward movement of production from Pennsylvania to the new centers along the Great Lakes, at Chicago, Cleveland, and Detroit.[1] The purpose of this paper is to present an econometric model of the industry's location; the results of estimating this model contradict the previous cost-related theories of the movement of iron and steel. I conclude that the locational shift occurred because of the differential growth of demand by steel-using industries; this growth itself took place for reasons independent of the steel industry.

The main characteristics of the westward movement of the industry can be seen in Table 1, which gives the shares of iron and steel production by states from 1910 to 1972. The three major producers—Pennsylvania, Ohio, and Illinois-Indiana (which represents almost entirely Chicago-Gary)—have remained the center of the industry. Growth of steel production has taken place mainly in centers which were already established in 1900;[2] it is the differential growth of these centers which has provided the change in the industry's primary location. Thus Pennsylvania's share has fallen from .51 to .23 over this century, while the share of Illinois-Indiana has risen from .15 to .25 and that of Michigan from .001 to .07. The model which is developed here incorporates the basic factors determining this change in the location of the industry. The parameters of the model are estimated using time-series data and the results are used to compare the cost-related and demand-related theories of the movement of production.

I. The Model

The essential feature of the market for iron and steel is the interaction of several major production centers competing in a national market. This means that the determination of output and price in any one production center is influenced by the structure of input prices at that location and by the demand for the product, which is not independent of the prices charged by producers located at other centers. Consumers will buy steel from the center offering the lowest delivered price. Therefore, the market region for a given steel center is determined by the geographical area in which that center's delivered price of steel is less than or equal to the prices offered by its competitors in other producing centers. In order to take account of this spatial interaction, the model is designed to measure the demand for steel in one production center as a function of the geographical pattern of steel prices and the location of demand.

A. *The Cost Function for Steel*

The supply of steel at one location is described by the cost function dual to the production function. It represents the aggregate supply conditions of all producers in a given center; each producer is assumed to be attempting to fulfill the conditions for profit maximization. The most general homogeneous function which has been developed to approximate production and

*Assistant professor of economics, Boston College. This paper is taken from my doctoral dissertation. I wish to acknowledge the help of my committee, George Tolley, Donald McCloskey and D. Gale Johnson; and the useful suggestions of R. E. Lucas, Jr. and Gideon Fishelson.

[1]See Gunner Alexandersson (1961, 1967); Walter Isard; Isard and W. Capron; and George Stocking.

[2]This is documented in Alexandersson (1961).

TABLE 1—INGOTS AND STEEL FOR CASTINGS, STATE DISTRIBUTION, 1910–72

State	Share of Total *U.S.* Production						
	1972	1960	1950	1940	1930	1920	1910
Illinois-Indiana	.251	.222	.201	.210	.206	.177	.154
Pennsylvania	.228	.240	.282	.303	.353	.419	.506
Ohio	.179	.173	.195	.206	.226	.239	.194
Subtotal	.658	.635	.678	.719	.785	.835	.854
Michigan	.070	.066	.057	.052	.012	.001	.001
Maryland	.047	.072	.059	.058	.037	.017	.014
New York	.031	.052	.049	.056	.043	.046	.051
California	.027	.027	.025	.012	.010	.006	.000
Total	.833	.852	.868	.897	.887	.905	.920

Source: *AISI*, *Annual Statistical Report*, 1912–72.

cost relationships is the "transcendental logarithmic" or "translog" function.[3] This allows cost to be expressed as

$$C = g(Q)f(p_1, p_2, \ldots, p_n) \tag{1}$$

where C is total cost, Q is output, and the p's are factor prices. The function $f(\cdot)$ is linear homogeneous in the factor prices. The output term $g(Q)$ represents the scale of output; it can be approximated by any general functional form. Thus the translog function is not constrained to a constant economies-of-scale factor as are both the *CES* and Cobb-Douglas. It is possible to have a U-shaped or L-shaped marginal cost curve by specifying $g(Q)$ correctly.

The industry supply curve for a region is obtained by the condition that marginal cost equals price:[4]

$$\frac{\partial C}{\partial Q} = P = g'(Q)f(p_1, \ldots, p_n) \tag{2}$$

The translog cost function is written in *log* form as

$$ln\,P = ln\,g'(Q) + v_0 + \sum v_i\,ln\,p_i + \frac{1}{2}\sum_i \sum_j \gamma_{ij}\,ln\,p_i\,ln\,p_j \tag{3}$$

where v_0, v_i, and γ_{ij} are parameters of the cost function. Equation (3) can also be regarded as a logarithmic Taylor series expansion to the second term around input prices of an arbitrary twice differentiable cost function.[5] The translog function incorporates variable partial elasticities of substitution. It has as special cases the *CES* and Cobb-Douglas functions (equation (3) is Cobb-Douglas when all $\gamma_{ij} = 0$). Equation (3) is the supply curve for the industry in one region, expressed in terms of all factor prices and physical output. The estimation of a supply relationship such as (3) for the steel industry allows the model to discriminate between the effects of location-specific factor prices and returns to scale in determining supply price. These elements of supply capture the basic differences between alternative production centers.

B. *Demand*

The demand for steel must be measured separately for each production center because of the effect of distance on quantity demanded. Consumers of steel are widely

[3]See Laurits R. Christensen et al. and Hans P. Binswanger.

[4]This does not preclude the possibility of imperfect competition in the steel industry. Since a constant elasticity demand function is used along with the supply function developed here, price and marginal revenue are related by a constant: $MR = P(1 + 1/\eta) = kP$. Thus when regression estimates are made using the *log* form of the constant elasticity demand relation, parameter estimates will be unaffected by the alternative $MR = MC$ or $P = MC$.

[5]See Binswanger, p. 966.

dispersed, so the delivered price inclusive of freight charges is unique to each consumer. Steel will be purchased from the producer who offers the lowest delivered price, *ceteris paribus*, and this serves to define a watershed market area for each producer given the structure of freight rates.

A change in the f.o.b. mill price of steel in one production center will have two separate effects on quantity demanded: first, there will be the normal reaction of consumers' demand to a change in price; second, a change in price at one production center while the price at all others remains constant will affect the size of the market area itself. Therefore, the price of steel in competing centers should be included in the demand function estimated for each center. This was not possible because of the high correlation between steel prices in neighboring regions. Thus the measured elasticity of demand also includes the effects of changes in the size of the market region, if any.

The demand for steel in each production center is approximated by a function of the form:

$$Q = BP^{-\eta}R^{\theta}V^{\phi} \tag{4}$$

where B is a constant, R is the price of the major substitute for steel, V is an index of products made with steel (from which the demand for steel is derived), and η, θ, and ϕ are the constant demand elasticities with respect to P, R, and V.

Distance from producer to consumer affects the delivered price of steel and thus influences demand; therefore, a change in the geographical distribution of demand will affect the quantity demanded. The demand index V must represent the entire market area of each production center. The proper approach would be to estimate a demand function for each point of consumption; since this is not possible, a method of aggregating V is needed. The technique adopted here uses a variant of the "gravity weights" common to the location economics literature. An observation of V is obtained from each Standard Metropolitan Statistical Area (*SMSA*) in the market area; these V's are weighted not by the reciprocal of distance squared, as is customary, but by $w_i = 1/(1 + c_i)$ where c_i is the transportation cost from producer to consumer, expressed as a percent of P. Since transportation cost is approximately linear with respect to distance with a fairly large constant term (or loading cost), the average cost of distance is declining. This means that the w_i give more weight to consumers located close to the production centers, but the weights decline with distance at a decreasing rate to reflect the structure of transportation costs.

The demand function (5) is estimated in *log* form:

$$\ln Q = \ln B - \eta \ln P + \theta \ln R + \phi \ln \left[\sum_i w_i V_i\right] \tag{5}$$

where $\sum_i w_i V_i = V$, the aggregated demand index.

II. Estimation of the Model

The two-equation model of regional supply and demand was fitted to data covering the period 1921-72; the three principal steel production centers—Pennsylvania, Ohio, and Chicago—were included. The variables are defined in Table 2. The estimated cost function includes the prices of five factor inputs—labor, capital, steel scrap, iron ore, and coke—which account for over 90 percent of the cost of producing steel. Aluminum is used as the substitute for steel in the demand function.

The demand shift variable, in accordance with equation (5), must be an index of the level of demand for steel in each consumption location. The industries which use steel are too numerous to be included separately in the demand function; only the auto and construction industries consume more than 6 percent of steel production. It was found that value-added in manufacturing (net of iron and steel) was a superior index of demand. While the relative importance of industries such as autos and railroads fluctuated over the century, the share of the value-added by iron and steel in total value-

TABLE 2—DEFINITIONS OF VARIBLES USED IN REGRESSION EQUATIONS

Variable	Definition
P	*Log* of weighted regional price index for steel products divided by the Wholesale Price Index (*WPI*)
Q	*Log* of tons of steel produced in the region
R	*Log* of the price of aluminum ingots divided by *WPI*
V	*Log* of value-added in manufacturing (less *VA* in iron and steel) by *SMSA* for each market area, divided by the transportation cost factor $(1 + c)$ and summed over all *SMSA*'s in the market region, divided by the *WPI*
W	*Log* of yearly wages in the iron and steel industry by region, divided by the *WPI*
K	*Log* of $P_k(r + d)/WPI$, where P_k = a price index of capital goods, r = Moody's Industrial Bond Rate, and d = rate of depreciation for capital goods in the steel industry.
S	*Log* of the price per ton of No. 1 Heavy Melting Steel Scrap in each region, divided by the *WPI*
C	*Log* of price per ton of blast furnace coke in each region, divided by the *WPI*
O	*Log* of the price of iron ore per ton, at the dock of lower Great Lakes ports, divided by *WPI*
$D1$	Dummy variable for Pennsylvania
$D2$	Dummy variable for Ohio

added by manufacturing remained quite stable over the period used for the estimation (this share was 3.0 percent in 1921 and 3.5 percent in 1972, with a range of variation of about 1.8 percent). Given the changing industry definitions for many steel-consuming industries over time and the changing steel requirements within each, total value-added is as accurate an index of demand as a disaggregated measure.

The model which was fitted to the data for 1921–72 consists of two equations estimated by two-stage least squares:

$$(6) \quad \ln Q = \ln B + \theta \ln R - \eta \ln P + \phi \ln\left[\sum_i w_i V_i\right] + \epsilon_1$$

$$\ln P = v_0 + \gamma_0 \ln Q + \sum_i v_i \ln p_i + \sum_i \sum_j \gamma_{ij} \ln p_i \ln p_j + \epsilon_2$$

where ϵ_1 and ϵ_2 are stochastic error terms.

In order to obtain the maximum number of observations to estimate the model, the three regional samples were estimated jointly in one pair of regressions by pooling all the data and adding dummy variables for the separate regions. In addition, the model was estimated separately for each of the three regions. The supply equation is quite cumbersome, containing sixteen independent variables. The F tests for both the individual and the joint contribution of the variables suggested that the five factor prices were highly significant but that the cross products were not. This result indicates that the first-order approximation or Cobb-Douglas form of the translog function is the most appropriate here. The results of the estimations are reported in Tables 3 and 4.

The picture presented by the regression is one of demand and supply relationships which are quite similar in the three production regions. Although they experienced widely different rates of growth of steel output, each region seems to face a demand schedule with about the same price, cross price, and shifter elasticities. The supply equations also show a marked similarity; the factor price coefficients are comparable, while the dummy variables in Table 3 suggest that there is only a slight difference in the levels of the cost functions between areas.

The principal shortcoming of these results is the large increasing returns to scale factor suggested by the negative coefficients estimated for Q (output). The alternative explanation is that the real price of steel decreased over time for other reasons as output increased. To test the specification of the model, two additional pieces of information were used. One is that technological change over the period was to a considerable degree embodied in new blast

TABLE 3—REGRESSION RESULTS FOR THE POOLED SAMPLE. 1921–72

Demand Equation: Dependent Variable = Q

P	R	V	$D1$	$D2$
−1.43	0.21	1.01	−.066	−.009
(−6.9)	(1.9)	(12.7)	(−.13)	(−.24)

$R^2 = .90$ $F = 177.4$

Supply Equation: Dependent Variable = P

W	K	S	O	C	$D1$	$D2$	Q
.44	.27	.10	.37	−.09	.04	−.03	−.19
(6.7)	(6.1)	(3.0)	(7.8)	(−3.6)	(2.8)	(−2.5)	(−4.8)

$R^2 = .93$ $F = 160.5$

Note: t-statistics are in parentheses beneath the estimated coefficients.

furnaces and steel mills which greatly enlarged the scale of output. A measure of this improvement in productivity would be the size of new steel plants. However, this data is not available after 1960, when the American Iron and Steel Institute (*AISI*) stopped publishing its annual census of steel plants. The available information which most closely approaches the desired data is the number of blast furnaces in existence. It can be obtained by region. The reasoning here is that for a given level of output, the fewer the number of blast furnaces the greater is the technological level, and thus the lower price will be.

The second enlargement of the model involves one of the location theories to be discussed in Section III. Basing-point pricing governed steel prices under the single basing point (Pittsburgh) until 1924 and under multiple basing points (including Pittsburgh, Cleveland, and Chicago) until 1948. The abolition of the system was alleged to have destroyed some monopoly power in the industry. To test for this effect, a dummy variable was entered for the years

TABLE 4—REGRESSION RESULTS FOR THE SEPARATE REGIONS, 1921–72

Demand Equation: Dependent Variable = Q

	P	R	V	R^2	$D.W.$	F
Chicago	−1.36	.19	1.00	.93	0.8	129.0
	(−3.1)	(.83)	(6.9)			
Pennsylvania	−1.66	.15	1.09	.84	1.6	52.7
	(−3.8)	(.64)	(5.6)			
Ohio	−1.33	.46	.98	.90	1.4	88.5
	(−4.9)	(2.5)	(8.3)			

Supply Equation: Dependent Variable = P

	W	K	S	O	C	Q	R^2	$D.W.$	F
Chicago	.68	.19	.21	.24	−.04	−.30	.91	1.6	45.2
	(2.7)	(1.6)	(2.0)	(1.5)	(.45)	(−2.3)			
Pennsylvania	.64	.28	.29	.18	−.12	−.38	.92	1.3	52.0
	(4.2)	(2.8)	(2.7)	(1.4)	(−2.2)	(−3.5)			
Ohio	.49	.26	.07	.31	−.09	−.33	.83	1.3	22.2
	(2.9)	(1.7)	(1.0)	(2.3)	(−.94)	(−3.0)			

Note: t-statistics shown in parentheses.

TABLE 5—REGRESSION RESULTS FOR THE EXPANDED MODEL, POOLED 1921-72

Demand Equation: Dependent Variable = Q

P	R	V
−1.37	.19	.99
(−7.8)	(2.0)	(15.4)

$R^2 = .90$ $F = 303.7$

Supply Equation: Dependent Variable = P

W	K	S	O	C	BF[a]	DM[b]	Q
.36	.22	.04	.43	.07	.06	.17	−.01
(8.7)	(6.7)	(2.0)	(13.7)	(3.4)	(3.0)	(9.8)	(−0.4)

$R^2 = .97$ $F = 364.1$

[a] BF = the *log* of the number of blast furnaces by region
[b] DM = dummy variable for the years prior to 1948

before 1948. A positive coefficient would suggest that industry prices were higher under basing-point pricing.

The parameter estimates of the expanded model are reported in Tables 5 and 6. The two new coefficients have the expected signs and are significant; the coefficient of Q in Table 5 is virtually zero, suggesting that the regional production functions for steel have constant returns to scale. The coefficient of the price of coke is now positive in all regions. It is this model, as amended in Tables 5 and 6, which will be used in Section III to test several possible causes for the changing location of steel.

The regression results which have been presented here suggest that the Cobb-Douglas production function fits the data

TABLE 6—REGRESSION RESULTS FOR THE EXPANDED MODEL, 1921-72

Demand Equation: Dependent Variable = Q

	P	R	V	R^2	D.W.[a]	F
Chicago	−1.27	.15	.97	.93	0.80	130.6
	(−3.0)	(.67)	(6.8)			
Pennsylvania	−1.62	.13	1.08	.84	1.55	53.0
	(−3.8)	(.58)	(5.5)			
Ohio	−1.21	.40	.93	.90	1.42	91.0
	(−4.6)	(2.2)	(8.2)			

Supply Equation: Dependent Variable = P

	W	K	S	O	C	BF	DM	Q	R^2	D.W.[a]	F
Chicago	.55	.18	.11	.32	.09	−.48	.14	−.10	.97	1.96	100.2
	(3.7)	(2.5)	(1.6)	(3.2)	(1.5)	(−1.6)	(3.4)	(−1.3)			
Pennsylvania	.47	.17	.10	.33	.07	.05	.14	−.14	.97	1.46	108.9
	(4.7)	(2.2)	(1.2)	(3.9)	(1.0)	(1.0)	(3.0)	(−1.6)			
Ohio	.55	.05	.02	.30	.23	.15	.30	−.08	.96	1.95	88.1
	(6.9)	(.66)	(.63)	(4.7)	(3.9)	(2.4)	(8.3)	(−1.4)			

Note: t-statistics shown in parentheses.

[a] Durbin-Watson statistics are in the indeterminate range. A more complete test of the time-series properties of the residuals was performed, using Box-Jenkins analysis. No significant autocorrelations or partial autocorrelations were found, using up to a 10-year lag. For a description of the procedure, see Charles Nelson.

TABLE 7—Factor Share Comparisons

Input Price	Cost Share	Regression Coefficient
Labor	.38	.36
Capital	.20	.22
Scrap	.15	.04
Ore	.10	.43
Coke	.10	.07
Total	.93	1.12

Sources: AISI, Annual Statistical Report; Iron Age; and Mineral Yearbook. The quantities of material inputs were multiplied by market prices and divided by total industry revenue to obtain factor shares; labor and capital shares were taken from the combined income statement of the industry. Regression coefficients are from Table 5.

well. An additional test of this relationship is that, with constant returns to scale, the cost shares of the factors of production should be equal to the estimated coefficients of the factor prices. A comparison of these cost shares and regression coefficients is given in Table 7. The correspondence is quite close between the two sets of estimates except for the coefficient of iron ore, which is four times as large as expected.

III. Previous Theories of Steel's Westward Movement

The economics of the location of the iron and steel industry has interested numerous writers over the last fifty years. The importance which these writers have placed on such factors as deposits of iron ore and coal is not challenged here. Rather, a question is raised regarding the reason for the movement of production from certain favorable locations to others. The differential growth of steel production by regions in the twentieth century is almost entirely explained by variation in the rates of growth of existing steel centers (see Table 1 and discussion). The central issue concerns what locational factors changed in those centers over the period.

The focus of the westward movement of iron and steel is the slow growth of the older Pennsylvania region relative to Chicago-Gary. Pennsylvania's average yearly rate of growth of output for 1921-72 was 2.21 percent, compared with 3.51 percent for Chicago. The excess of Chicago's rate of growth over that of Pennsylvania, 1.30 percent per year, represents the regional shift which must be explained. There are two often-cited theories of this movement of production; one is related to changing fuel requirements over time; and the other emphasizes the effects of the basing-point pricing system. The model which was estimated in the previous section is used to analyze these theories. Then the demand-related explanation is presented.

A. *The Distance-From-Coal Hypothesis*

In attempting to predict the future locational pattern of iron and steel, Isard concentrates on the effect of significant reductions in fuel consumption through the use of higher pressures in the blast furnace and of oxygen and enriched air in steel furnaces. Isard believes that these developments "... like other measures aimed at fuel economy in the past, will reduce transport expenditures on coal for distance centers ... by a greater *absolute* amount than for centers close by coal sites. Hence they will affect the latter adversely and benefit the former" (pp. 124 25).

The implications of Isard's theory can be seen more clearly by formulating it in terms of demand and supply. Improvements in the use of coal to make steel can be viewed as the result of changes in the parameters of the cost function (equation (3)). The effects of these changes on the cost functions in Chicago and Pennsylvania depend on the extent of the technical change and on the regional difference in the price of inputs.

The form of the cost function which was used in the regression estimation is

$$(7) \quad \ln P = v_0 + \sum v_i \ln p_i + \gamma_0 \ln Q + \delta_1 \ln BF + \delta_2 DM$$

Differentiating (7) totally with respect to time and allowing the v_i to change,

(8)
$$\frac{d\ln P}{dt} = \frac{dv_0}{dt} + \sum_i v_i \frac{d\ln p_i}{dt} + \sum_i \ln p_i \frac{dv_i}{dt} + \gamma_0 \frac{d\ln Q}{dt} + \delta_1 \frac{d\ln BF}{dt} + \delta_2 \frac{d\,DM}{dt}$$

The dv_i/dt are the changes in factor shares over time. Improvements in steelmaking reduced the use of coal (and therefore coke). The contribution of coke to changes in the supply price of steel consists of two parts:

$$v_c \frac{d\ln p_c}{dt} + \ln p_c \frac{dv_c}{dt}$$

where the subscript c refers to coke. Isard's point is concerned with a net advantage of Chicago over Pennsylvania due to these changes. This net advantage, which will be termed π, depends on the coke price differential between the two regions:

$$(9) \quad \pi = v_c\left(\frac{d\ln p_{c2}}{dt} - \frac{d\ln p_{c1}}{dt}\right) + (\ln p_{c2} - \ln p_{c1})\frac{dv_c}{dt}$$

Here a 1 refers to Pennsylvania and 2 to Chicago; π can be estimated from data on coke prices and the behavior of the share of coke over time. Data on the share of coke are spotty and somewhat inconsistent over time. Table 8 presents estimates of this share from 1921–72; it shows a rather steady downward trend from .12 to .07, although this is contradicted by the regression results reported earlier. Table 8 is inclusive because the method of calculation of coke's share changed over time, of necessity, since

TABLE 8—THE COST SHARE OF COKE

Year	Estimated Share	Year	Estimated Share
1921	.12	1950	.10
1925	.11	1955	.09
1929	.09	1960	.07
1935	.09	1965	.05
1940	.10	1970	.07
1946	.10	1972	.07

Sources: AISI, Annual Statistical Report and *U.S. Mineral Yearbook*, indicated years.

the form of the data changed (the regression equations do not suffer from this problem, as they rely on consistent data series). However, in order to bias the results in favor of the Isard hypothesis, the information contained in Table 8 is used along with the levels and rates of change of coke prices to quantify equation (9):

$$\pi = (.10)(-1.74) + (.224)(-.008) = -.176$$

This represents the yearly fall in the price of steel in Chicago relative to Pennsylvania. Multiplying π by the elasticity of demand in Chicago (from Table 6), the effect on output is $(-1.76)(-1.27) = .224$ percent per year. The dominant factor here is the first term of (9), which results from the steeper rise in the price of coke in Pennsylvania over the period. The effect emphasized by Isard, the coke-saving technology, accounts for very little by this measure. Both factors together explain $0.224/1.3 = 17.2$ percent of the excess of Chicago's growth rate over that of Pennsylvania. The technology factor alone explains $(.224)(-.008)(-1.27)/1.3 = .18$ percent of the shift.

B. *Basing-Point Pricing*

The other major theory of steel's location to be considered here is that put forward by George W. Stocking. He believed that the "Pittsburgh Plus" system of steel pricing which reigned from 1906 to 1924 retarded the development of all steel centers except Pittsburgh. This pricing system forced up the selling price at steel centers outside Pittsburgh by an amount equal to the freight rate from Pittsburgh to the given center. When the single basing-point method was prohibited in 1924, Chicago became a basing point. If the price of steel f.o.b. Chicago had been kept artifically high by Pittsburgh Plus, then a fall in the Chicago price by the amount of Pittsburgh-Chicago freight rates would be expected.

The regression results reported in Tables 5 and 6 show a higher price of steel in all three regions during the multiple basing-point system (1924–1948) than afterward. The evidence for this is the positive coefficient of the dummy variable, DM. This

suggests that Stocking may have been right about the effect of the basing-point system on prices. In order for this to have a locational effect, however, the end of the Pittsburgh Plus system in 1924 must have caused a fall in steel prices relative to Pittsburgh, stimulating demand in Chicago. These price changes are generally supported by the available data; the prices of steel products were about 14 percent higher in Chicago than in Pittsburgh during the reign of Pittsburgh Plus (1906–24), but they were essentially equal by the mid-1920's. This price differential corresponds closely to the freight cost between the two cities, which averaged 10–15 percent.[6] This correspondence could be due to a rigid adherence to the basing-point system or to competitive forces, if Pittsburgh were a low cost producer and forced Chicago to meet Pittsburgh's price in the Chicago market. That is, critics of basing-point pricing assume that Chicago's price is artifically held up to the price f.o.b. Pittsburgh plus freight to Chicago, while it is just as possible that the price of steel made in Chicago was kept down to Pittsburgh Plus by competition from Pittsburgh. The alternative to the basing-point theory of the fall in Chicago's price is that input prices in Chicago fell over the period. This is supported by the data on factor prices; wages and scrap prices rose less in Chicago than in Pittsburgh during the 1920's, and coke prices fell by a greater amount in Chicago, while ore and capital prices presumably behaved identically in both areas.

It will be assumed for the purpose of testing the Stocking thesis that the elimination of Pittsburgh Plus was responsible for the fall in Chicago's price. The effect on the long-run output of steel in Chicago is then the change in price, −14 percent, times the elasticity of demand, −1.27, or 17.8 percent. This compares with a differential rate of growth in Chicago over Pennsylvania of 66.3 percent over the fifty-year period 1921–72 (the differential is greater for Pittsburgh versus Chicago since eastern Pennsylvania increased somewhat faster over the period than Pittsburgh). The fall in Chicago's price "explains" 26.8 percent of the increase in Chicago's output relative to Pennsylvania. And it does so on the tenuous assumption that all of that price change came about as a result of the abolition of Pittsburgh Plus. In the next section, a more complete explanation of the increase in Chicago's output is provided.

IV. The Growth of the Chicago Market for Steel

From 1910 to 1972, Chicago's share of all steel produced in the United States rose from 15 to 25 percent. Buffalo, Detroit, and Cleveland, all Great Lake cities, grew in steel output relative to Youngstown and Pittsburgh. This change in steel's location has been characterized by many writers mainly by its direction of movement—west. The economic theories constructed to explain this growth have been, for the most part, theories of cost advantage. For reasons of natural resources (Isard) or the elimination of anticompetitive practices (Stocking), the movement of the industry has been linked to geographical differences in the cost of producing steel.

The cost-related theories do not stand up primarily because the cost changes which occurred in the twentieth century are insufficient to explain the large shift of production. The estimated supply curve is horizontal, so the only way for cost changes to induce increases in output is through a downward shift of the cost curve and a consequent increase in demand. The estimated demand curve in Chicago is not elastic enough to explain output growth in this way, given the modest change in relative steel prices between the old and new areas.

The model which has been estimated provides parameter estimates which can be combined with rates of change in the variables to obtain the relative contribution of each variable to output growth in Chicago over Pennsylvania. Table 9 shows the comparison of these contributions. Columns 3 and 6 contain the product of the average yearly percent change in the variables and

[6]Price differences and freight charges are based on weekly quotations from *Iron Age*, 1900–24.

TABLE 9—CONTRIBUTIONS OF DEMAND VARIABLES TO CHANGES IN OUTPUT, 1921–72

Variable	Chicago: Average Change[a] (1)	Estimated Coefficient[b] (2)	Columns (1) x (2) (3)	Pennsylvania: Average Change (4)	Estimated Coefficient (5)	Columns (4) x (5) (6)	Columns (3) – (6) (7)
Q	3.51				2.21		
V	4.88	.97	4.74	3.48	1.08	3.75	.99
R	–1.74	.15	–.26	–1.74	.13	–.23	–.04
P	0.78	–1.27	–.98	0.78	–1.62	–1.26	.27

Note: Columns (1) and (4) are shown in percent.
[a] *Source:* Regression of *log* of each variable on time.
[b] *Source:* See Table 6.

the estimated coefficients from the model; the result is the contribution of each variable to the excess of Chicago's growth rate over that of Pennsylvania.

The amount of the 1.3 percent yearly growth of Chicago relative to Pennsylvania explained by the demand index *V* is .99 percent, or 76 percent of that excess growth rate. This in turn is principally due to the difference in the rate of growth of demand since the other component of the contribution, the estimated coefficient, differs only slightly between the two centers (and it differs in favor of Pennsylvania's growth). The difference in cross elasticities of demand is quite small and results in a slight reduction in Chicago's relative growth. The average yearly growth of the price of steel, approximately equal in both areas, serves to increase demand in Chicago relative to Pennsylvania because of the somewhat more inelastic demand curve estimated for Chicago.

The factors which served to increase the supply price of steel in the two areas can be compared in the same way as the demand variables. Table 10 presents these contributions (columns 3 and 6) along with the differential changes between the two areas (column 7). There is no dominant variable here. Wages and capital costs contributed the most to price increases, but no significant regional differences appear.

The main conclusion to be drawn from the comparison of Chicago and Pennsylvania over the fifty-year period is that demand is the dominant factor in explaining Chicago's higher rate of growth. If demand

TABLE 10—CONTRIBUTIONS OF SUPPLY VARIABLES TO CHANGES IN PRICE, 1921–72

Variable	Chicago: Average Change[a] (1)	Estimated Coefficient[b] (2)	Columns (1) x (2) (3)	Pennsylvania: Average Change (4)	Estimated Coefficient (5)	Columns (4) x (5) (6)	Columns (3) – (6) (7)
P	0.78			0.78			
W	2.19	.55	1.20	2.32	.47	1.10	0.11
K	1.40	.18	0.25	1.40	.17	0.23	0.02
S	0.36	.11	0.39	–.01	.10	–.001	0.04
O	0.10	.32	0.03	0.10	.33	0.03	–.001
C	0.58	.09	0.05	2.31	.07	0.16	–.10
BF	0.13	–.48	–.06	–1.51	.05	–.07	.006
DM	–2.65	.14	–.38	–2.65	.14	–.37	–.005
Q	3.51	–.10	–.37	2.21	–.14	–.31	–.06

Note: Columns (1) and (4) are shown in percent.
[a] *Source:* Regression of *log* of each variable on time.
[b] *Source:* See Table 6.

had increased at an equal rate in both regions, quantity produced would have grown at equal rates also, because the supply and demand elasticities differ only slightly. The analysis presented in Section III regarding the cost-related theories has attempted to show that neither the historical data on cost changes nor the results of the estimated model supports the idea that changes in regional cost advantage explain the shifting pattern of steel output.

V. Conclusion

This paper has compared alternative (but not mutually exclusive) theories of the location of iron and steel production, using the results of an econometric model. The two previous explanations of the movement of the industry are cost related. The econometric evidence, on the other hand, indicates that cost differences had little or nothing to do with this movement. This is not to say that the location of production is not ultimately determined by cost considerations. In fact, all three production centers studied here became major steel producers because they combined low input prices with good distribution to steel markets. The issue is, what caused the relatively greater growth of the western producers (mainly Chicago-Gary)? And the answer is that the crucial variable in the twentieth century has been the differential growth of demand by steel-using industries.

REFERENCES

Gunner Alexandersson, "Changes in the Location Pattern of the Anglo-American Steel Industry: 1948–1959," *Econ. Geog.*, Apr. 1961, *37*, 95–114.

———, *Geography of Manufacturing*, Englewood Cliffs 1967.

H. P. Binswanger, "The Measurement of Technical Change Biases with Many Factors of Production," *Amer. Econ. Rev.*, Dec. 1974, *64*, 964–76.

L. R. Christensen, D. W. Jorgensen, and L. J. Lau, "Transcendental Logarithmic Production Frontiers," *Rev. Econ. Statist.*, Feb. 1973, *55*, 28–45.

J. S. Hekman, "An Analysis of the Changing Location of Iron and Steel Production in the Twentieth Century," unpublished doctoral dissertation, Univ. Chicago 1976.

W. Isard, "Some Locational Factors in the Iron and Steel Industry since the Early Nineteenth Century," *J. Polit. Econ.*, June 1948, *56*, 203–17.

——— **and W. Capron**, "The Future Locational Pattern of Iron and Steel Production in the United States," *J. Polit. Econ.*, Apr. 1949, *57*, 118–33.

Charles Nelson, *Applied Time Series Analysis for Managerial Forecasting*, San Francisco 1973.

George Stocking, *Basing-Point Pricing and Regional Development*, Chapel Hill 1954.

American Iron and Steel Institute (*AISI*), *Annual Statistical Report*, New York, 1912–72.

The Iron Age, New York, various issues.

U.S. Bureau of the Census, *Census of Manufacturers*, Washington 1921–72.

U.S. Bureau of Mines, *Mineral Yearbook*, Washington, various issues.

Internal Revenue Service, *Bull. F*, *Depreciation Guidelines*, Washington 1962.

The Impact of Demand and Price Expectations on the Behavior of Prices

By LOUIS J. MACCINI*

Empirical research on the behavior of prices has grown substantially in recent years. While a great deal has been learned about the factors primarily responsible for changes in prices, a number of issues either remain unresolved or have not yet been extensively analyzed. The purpose of this paper is to undertake an analysis of two of these issues: the influence of demand factors and of price expectations on prices.

William Nordhaus observed in a recent survey that empirical work on prices has failed to uncover a strong, systematic, and uniform influence of demand on prices. Some studies have found a relatively weak influence of demand on prices. Where a statistically significant effect has been found, the particular variable which captures the effect has varied widely from study to study, from sector to sector, and industry to industry even within the same study, and from sample period to sample period. This is quite unlike the impact of cost-push factors on prices where "normal" unit labor costs and some measure of raw material prices have consistently played an important role in explaining movements in prices.[1]

The empirical work that has produced these results is dominated by the use of so-called markup models that presume prices are set as a markup over unit factor costs. The influence of demand is introduced by assuming the markup is affected by "demand-pressure" factors. A large number of variables—including capacity utilization, unfilled order-shipment ratios, inventory-sales ratios, unfilled order-capacity ratios, etc.—has been examined in an effort to "pick up" the demand forces that impinge on prices.

There are several possible reasons for the failure of this approach to uncover a strong and systematic effect of demand on prices. First, the theoretical rationale for selecting this or that demand-pressure factor is often quite vague and imprecise. This not only raises questions about whether the theoretically appropriate demand factors have in fact been selected and whether their influence on prices has been properly specified, but also makes it difficult to interpret empirical evidence and to compare the effect of demand on prices with the effects of other factors. Second, the

*The Johns Hopkins University. I am indebted to Win Ogden, Robert Rossana, Andre Sapir, Jean Small, and Lisa Skumatz for research assistance and to the participants of the General Seminar at the Johns Hopkins University for helpful comments. I am also grateful to David Belsley for providing me with some unpublished data. Finally, I acknowledge with gratitude the financial support of the National Science Foundation under Grant # SOC 75-19653.

[1]For the sake of brevity, no attempt will be made here to survey the literature on price formation; extensive surveys have been done by Nordhaus as well as by Paul Earl, and by David Laidler and Michael Parkin. A recent paper by Robert J. Gordon, however, should be mentioned. Unlike the conventional literature he finds a strong effect of demand on prices by allowing for separate lags on the "demand variables." But, in this study too, uniformity across sectors is lacking. Furthermore, the demand variables that he selected are different from those used below. The price equations tested here are generalized versions of a price equation I used in a previous study of price and output behavior (1977). They are more general in two ways. First, the earlier model was constructed under some restrictive assumptions, namely, that industry demand is always completely price inelastic, that raw material prices have no effect on prices, and that expected inflation of the industry price level equals expected money wage inflation minus productivity. Each of these assumptions is relaxed in this paper. Second, in the empirical work of this paper, unlike previous work, I allow for different and separate lags on the effect of various factors on prices. The model is tested with a larger and more varied set of industries, and a more extensive test of the effect of price expectations on prices is made. These extensions lead to substantially different empirical conclusions about the determinants of prices.

treatment of lags in the effect of demand on prices is deficient. Some studies do not allow for a lag at all. Others, as Gordon has pointed out, that do allow for a lag typically do so by introducing the lagged dependent variable as an explanatory variable in the relevant price equation. This constrains the lag effect to be of the same form as that for other factors and to be one with geometrically declining weights. These restrictions are somewhat implausible, and to the extent that they are false, specification errors will arise in the empirical results. In any case, before the issue of the quantitative impact of demand on prices is resolved, it would seem that further analysis with a model that is free of these difficulties is clearly warranted.

A second issue to be investigated is the impact of price expectations on prices. Price expectations are often assumed to influence wages and thus, indirectly, prices.[2] That is, wage equations typically contain a term representing the expected rate of inflation of consumption goods' prices. The purpose is to allow for the notion that suppliers of labor take into account expectations concerning the cost of living when they make wage demands. But, as several authors have indicated in recent theoretical work,[3] price expectations may influence prices in another manner. If in making price decisions firms take into account the prices that their competitors are charging both now and in the future, then their expectations will directly affect the prices that they set. This factor would constitute a *direct* price-expectations effect on prices over and above any indirect effect operating through wages. An extensive empirical analysis of such an effect, however, has not yet been undertaken.

The next section of the paper constructs the theoretical framework. Section II reports the results of empirical work undertaken with data from the manufacturing sector of the *U.S.* economy, and Section III concludes the paper.

[2]See, for example, Otto Eckstein and Roger Brinner.

[3]See, for example, Ray Fair, Edmund Phelps and Sidney Winter, Hugh Rose, and the author (1977).

I. The Theoretical Framework

The basic theoretical framework assumes that the typical industry is composed of two different types of firms—those that produce to stock and those that produce to order.[4] This distinction is made to capture the separate roles that finished goods inventories and unfilled orders play in the process of price determination. An optimization model for each type of firm is developed, and each of these models yields an optimal price decision rule. These rules then serve as a basis for constructing an aggregate price behavioral relationship for the industry as a whole.

Space limitations, however, preclude a detailed analysis of the models of firm behavior and the aggregation process.[5] Hence, I present here an outline of the model for each type of firm and a statement of the aggregate price behavioral relationships that are to be used below in the empirical work.

A. *Production to Stock*

The optimization model for a firm that produces to stock is to maximize J_s with respect to $\{p^s(\tau), x^s(\tau), n^s(\tau), h^s(\tau): \tau \geq t\}$, where

$$(1) \quad J_s = \int_t^\infty f^s(\tau) \exp[-r_t^{es}(\tau - t)] d\tau$$

subject to

$$(2a) \quad f^s(\tau) = p^s(\tau) n^s(\tau) - c^s(x^s(\tau), h^s(\tau), w_t^{es}(\tau), m_t^{es}(\tau), a_t^{es}(\tau))$$

$$(2b) \quad c_1^s > 0, \quad c_2^s > 0, \quad c_3^s > 0, \quad c_4^s > 0, \quad c_5^s < 0$$

$$(3) \quad \frac{dh^s(\tau)}{d\tau} = x^s(\tau) - n^s(\tau) \qquad h^s(t) = h_t^s$$

$$(4a) \quad n^s(\tau) = n^s(p^s(\tau)/v_t^s(\tau), h^s(\tau)/q_t^{es}(\tau), q_t^{es}(\tau))$$

[4]See David Belsley for a discussion of the need to distinguish between production to stock and production to order.

[5]A complete description and analysis of the models of firm behavior are available from the author upon request.

(4b) $n_1^s < 0 \quad n_2^s > 0 \quad n_3^s > 0$

where $p^s(\tau)$ is the price that the firm sets, $x^s(\tau)$ is its rate of output, $n^s(\tau)$ is the inflow of new orders that it expects to receive and to be able to service (i.e., expected shipments), $h^s(\tau)$ is its stock of finished goods inventories, $f^s(\tau)$ is its expected net receipts, r_t^{es} is its expected interest rate, $w_t^{es}(\tau)$ is its expected wage rate, $m_t^{es}(\tau)$ is its expected raw materials price, $a_t^{es}(\tau)$ is its estimated index of technology, $v_t^s(\tau)$ is its estimate of the average price level prevailing in the industry, $q_t^{es}(\tau)$ is its estimate of the level of industry orders. The superscript s denotes that a variable pertains to a firm that produces to stock, t refers to calendar time, and τ denotes planning time.

The major characteristics of the model are the following:

(i) The firm explicitly sets prices in accordance with present value maximization principles. This aids in providing a precise theoretical underpinning for the various factors—demand, cost, or otherwise—that influence prices.

(ii) The firm is assumed to possess some monopoly power. Accordingly, the firm's demand conditions, (4a) and (4b), embody the assumption that the firm is able to control the flow of orders that it expects to receive and service by varying its "relative price," that is, by varying its price relative to the ruling average price level for stock-produced goods.[6] This relative price variable incorporates into the model the notion that the demand for the typical firm's product depends on its "competitive position" in the market. Because of imperfect information, however, each firm must estimate or forecast the relevant average price level both currently and in the future to make decisions, and consequently price expectations are built into the model.

(iii) Finished goods inventories enter the model in two ways. First, inventories appear in the firm's demand function, (4a), essentially to provide a motivation for the firm to hold inventories. Firms hold inventories primarily to protect themselves against "stockouts." We are assuming that as a first approximation, the larger is the firm's stock of inventories relative to its estimate of expected demand, the smaller is the possibility that the firm will be caught out of stock. Hence the higher it can expect its inflow of new orders, shipments, and sales to be.[7] Secondly, inventories also appear in the firm's cost function, (2a), to take account of inventory holding costs in the form of storage costs, insurance costs, and the like.[8]

(iv) Two alternative assumptions about the price elasticity of industry demand will be employed in the model. One is to assume that expected industry demand is completely price inelastic in which case the expected level of industry orders for stock-produced goods is the appropriate industry demand variable, and (4a) and (4b) then constitute a complete specification of the firm's demand conditions. An alternative is to allow expected industry demand to have some price elasticity in which case (4a) and (4b) need to be supplemented by an expected industry demand function which is assumed to take the following form:

(5a) $q_t^{es}(\tau) = q^{es}(v_t^s(\tau)/g_t^{es}(\tau), y_t^{es}(\tau))$

(5b) $q_1^{es} < 0 \quad q_2^{es} > 0$

where $g_t^{es}(\tau)$ is some expected general price level and $y_t^{es}(\tau)$ is an appropriate "shift" variable (for example, expected real income)

[6]Note that we are assuming that the firm when it alters its price believes that its actions will not alter the industry average.

[7]A more rigorous approach to inventory holding behavior by the firm would of course be to explicitly allow for uncertainty in the firm's optimization process. Due to the mathematical difficulties associated with dealing explicitly with uncertainty in intertemporal optimization models, I resorted to the above approach which, while an approximation, has the advantage of yielding price behavioral relationships together with predictions that may be subjected to empirical analysis.

[8]The cost function is essentially a short-run cost function with labor and raw materials as the variable factors of production. It is implicitly assumed that any fixed factors of production are either constant or are expected to grow at a constant rate and are incorporated into $a_t^e(\tau)$. This is a restrictive assumption and an interesting extension of the above model would be to integrate the price decision with decisions on investment in fixed factors of production.

which is constructed in such a way that an increase in $y_t^{es}(\tau)$ raises $q_t^{es}(\tau)$.

(v) To derive optimal decision rules for price, additional assumptions are necessary. In particular, a specific assumption must be made about the variables that the firm must estimate or forecast to make decisions.[9] We assume that the relevant expected magnitudes, namely, $v_t(\tau)$, $w_t^e(\tau)$, $m_t^e(\tau)$, $a_t^e(\tau)$, and $q_t^e(\tau)$ (or $g_t^e(\tau)$ and $y_t^e(\tau)$), are expected to change at constant exponential rates.[10] This assumption may be interpreted as stating that the firm is making price decisions on the basis of "normal" or "trend" levels of variables that it must forecast to make decisions. It is important to note that this assumption is quite consistent with the empirical literature on price formation. The latter has stressed that prices are determined by the normal levels of relevant explanatory variables, and the above assumption is one way of rigorously incorporating this notion into an intertemporal model of firm behavior. Using this terminology, the above assumption essentially means that at calendar date t, the firm must estimate both the current normal level and the normal growth rate of each of the expected magnitudes. How these estimates are related to observable variables will be specified below.

B. *Production to Order*

The optimization model for a firm that produces to order is to maximize J_o with respect to $\{p^o(\tau), x^o(\tau), n^o(\tau), u^o(\tau):\tau \leq t\}$, where

$$(6)\quad J_o = \int_t^\infty f^o(\tau)\exp[-r_t^{eo}(\tau - t)]\,d\tau$$

subject to

$$(7a)\quad f^o(\tau) = p^o(\tau)n^o(\tau) - c^o(x^o(\tau), u^o(\tau), w_t^{eo}(\tau), m_t^{eo}(\tau), a_t^{eo}(\tau))$$

$$(7b)\quad c_1^o > 0 \quad c_2^o < 0 \quad c_3^o > 0 \quad c_4^o > 0 \quad c_5^o < 0$$

$$(8)\quad \frac{du^o(\tau)}{d\tau} = n^o(\tau) - x^o(\tau) \qquad u^o(t) = u_t^o$$

$$(9a)\quad n^o(\tau) = n^o(p^o(\tau)/v_t^o(\tau), u^o(\tau)/x^o(\tau), q_t^{eo}(\tau))$$

$$(9b)\quad n_1^o < 0 \quad n_2^o < 0 \quad n_3^o > 0$$

when expected industry demand is assumed to be completely price inelastic. When some price elasticity to industry demand is allowed, the constraints need to be supplemented with the following industry demand function:

$$(10a)\quad q_t^{eo}(\tau) = q^{eo}(v_t^o(\tau)/g_t^{eo}(\tau), y_t^{eo}(\tau))$$

$$(10b)\quad q_1^{eo} < 0 \qquad q_2^{eo} > 0$$

Here $u^o(\tau)$ is the firm's stock of unfilled orders, and all other symbols are defined as above where the superscript *o* indicates that the variables pertain to a firm that produces to order.

The major characteristics of the model of an order-producing firm are essentially the same as that of a stock-producing firm except that the stock of unfilled orders rather than finished goods inventories is the primary concern of the firm. The stock of unfilled orders enters the firm's optimization process in two main ways. First, it enters the cost function to capture the notion that larger order backlogs enable the firm to achieve cost savings by combining orders for production which reduces set-up costs and permits the scheduling of smoother production flows. Secondly, the stock of unfilled orders, when expressed as a ratio to the rate of output, is an approximate measure of the firm's delivery lag and as such it enters the firm's demand function on the grounds that longer lead times will reduce the inflow of new orders to the firm.

C. *Price Behavioral Relationships*

Each of the above models yields an optimal price decision rule for $\tau \geq t$. Each firm is assumed to carry out its plan for calendar date t, based on its estimates of market

[9]Other assumptions are that the cost function and demand function are assumed to be log-linear and that the decision rules are computed by considering a linear approximation of the firm's optimality conditions about an appropriate equilibrium.

[10]An exception is the interest rate which is expected to remain constant at its normal level.

parameters and its stock of inventories or unfilled orders at that date. Assuming further that each firm continuously revises its plans and repeats the above process, we may derive a price behavioral relationship for each type of firm that describes the actual behavior of prices at any date t of calendar time.

These price behavioral relationships, however, are for individual firms. If empirical work is to be undertaken with time-series data for industrial aggregates, then some aggregation of these relationships is inevitably necessary. Since we have assumed that industries are composed of two different types of firms, those that produce to stock and those that produce to order, it is natural to aggregate in two stages. In the first stage, aggregate relationships for each type of firm would be derived, and in the second stage, the latter would be combined into an aggregate relationship for the industry as a whole. To avoid a lengthy derivation of such an aggregation process, I merely state the aggregate price behavioral relationships that emerge from such a process. They will be used below in the empirical work.[11]

The model yields two aggregate behavioral relationships for prices depending on what assumptions are made about the price elasticity of industry demand. If expected industry demand is completely price inelastic, then the aggregate price behavioral relationship is

$$(11a)\quad \ln P_t = \alpha_0 + \alpha_1 \ln H_t + \alpha_2 \ln U_t + \alpha_3 \ln Q_t^e + \alpha_4 \ln(W_t^e/V_t) + \alpha_5 \ln(M_t^e/V_t) + \alpha_6 \Phi_t + \alpha_7 \ln A_t^e + \ln V_t$$

$$(11b)\quad \alpha_1 < 0 \quad \alpha_2 > 0 \quad \alpha_3 > 0 \quad 0 < \alpha_4 < 1 \quad 0 < \alpha_5 < 1 \quad \alpha_6 < 0 \quad \alpha_7 < 0$$

[11]It should be stressed that due to data limitations and mathematical difficulties the derivation of such aggregate relationships will require some rather restrictive assumptions and approximations. See H. A. John Green and Lawrence Klein for a discussion of the kinds of assumptions and approximations that are needed to construct aggregate relationships of the form that are used here.

where $\Phi_t = R_t^e - d \ln V_t/dt$. Alternatively, if expected industry demand is allowed to have some price elasticity, then it is necessary to specify the following industry demand function:

$$(12a)\quad \ln Q_t^e = \eta^o + \eta_1 \ln(V_t/G_t^e) + \eta_2 \ln Y_t^e$$

$$(12b)\quad \eta_1 < 0 \quad \eta_2 > 0$$

Substituting (12a) into (11a), we may derive an alternative aggregate price behavioral relationship, namely,

$$(13a)\quad \ln P_t = \hat{\alpha}_0 + \alpha_1 \ln H_t + \alpha_2 \ln U_t + \gamma_1 \ln(V_t/G_t^e) + \gamma_2 \ln Y_t^e + \alpha_4 \ln(W_t^e/V_t) + \alpha_5 \ln(M_t^e/V_t) + \alpha_6 \Phi_t + \alpha_7 \ln A_t^e + \ln V_t$$

$$(13b)\quad \hat{\alpha}_0 = \alpha_0 + \alpha_3\eta_0 \quad \alpha_1 < 0 \quad \alpha_2 > 0 \quad \gamma_1 = \alpha_3\eta_1 < 0 \quad \gamma_2 = \alpha_3\eta_2 > 0 \quad 0 < \alpha_4 < 1 \quad 0 < \alpha_5 < 1 \quad \alpha_6 < 0 \quad \alpha_7 < 0$$

In these equations, P_t is the actual average price level, V_t is the expected normal average price level, H_t is the stock of finished goods inventories, U_t is the stock of unfilled orders, Q_t^e is the expected normal level of new orders in the industry, W_t^e is the expected normal money wage rate, M_t^e is the expected normal price of raw materials, A_t^e is an index of the expected level of technology, G_t^e is the expected normal level of a general or economy-wide price level, Y_t^e is the expected normal level of an appropriate "shift" variable, say, real national income, Φ_t is the expected normal real rate of interest, and R_t^e is the expected normal nominal rate of interest. Upper case letters represent appropriate averages or aggregates of individual firm variables which were denoted above by corresponding lower case letters.

The model provides a well-defined specification of the various factors that theoretically should influence prices. These include not only the familiar "cost-oriented" factors, such as normal wage rates, normal raw material prices, and the level of technology (the growth of which represents "productivity"), but also a clear and precise set of "demand-oriented" factors. The demand-oriented factors are of two types: First, the expected normal level of industry

demand is a demand-oriented factor that influences prices. If expected industry demand is completely price inelastic, then, as in (11a), the expected normal level of new orders in the industry is an adequate specification of expected demand considerations. But, if expected industry demand possesses some price elasticity, then, as in (13a), the determinants of the expected normal level of industry orders need to be specified and included in the model. Secondly, the stocks of finished goods inventories and unfilled orders may also be interpreted as demand-oriented factors that influence prices in that they essentially act as buffers that absorb unexpected changes in demand. The empirical work will attempt to isolate the quantitative importance of these demand-oriented factors in explaining movements in prices.

A second feature is that in making price decisions firms are assumed to take into account the prices that they believe their competitors are charging both currently and in the future. This implies that price expectations will have a direct impact on prices over and above any indirect impact operating through wage rates. Price expectations appear in the form of the current expected normal average price level and the expected rate of inflation of the normal price level in (11a) and (13a).

II. Empirical Work

A. *The Empirical Model*

To obtain the empirical price equations, we convert (11a) and (13a) to the appropriate discrete time analogues, take first differences of the resulting equations, and add error terms. When expected industry demand is completely price inelastic, the relevant price equation is

$$\text{(14a)}\quad \Delta\, ln\, P_t = \alpha_1 \Delta\, ln\, H_{t-1} + \alpha_2 \Delta\, ln\, U_{t-1} + \alpha_3 \Delta\, ln\, Q_t^e + \alpha_4 \Delta\, ln\,(W_t^e/V_t) + \alpha_5 \Delta\, ln\,(M_t^e/V_t) + \alpha_6 \Delta\Phi_t + \alpha_7 \Delta\, ln\, A_t^e + \alpha_8 \Delta\, ln\, V_t + \epsilon_{1t}$$

$$\text{(14b)}\quad \alpha_1 < 0 \quad \alpha_2 > 0 \quad \alpha_3 > 0 \quad 0 < \alpha_4 < 1 \quad 0 < \alpha_5 < 1 \quad \alpha_6 < 0 \quad \alpha_7 < 0 \quad \alpha_8 = 1$$

where $\Delta\Phi_t = \Delta R_t^e - \Delta\, ln\, V_t + \Delta\, ln\, V_{t-1}$. Alternatively, when expected industry demand is assumed to have some price elasticity, the appropriate price equation is

$$\text{(15a)}\quad \Delta\, ln\, P_t = \alpha_1 \Delta\, ln\, H_{t-1} + \alpha_2 \Delta\, ln\, U_{t-1} + \gamma_1 \Delta\, ln\,(V_t/G_t^e) + \gamma_2 \Delta\, ln\, Y_t^e + \alpha_4 \Delta\, ln\,(W_t^e/V_t) + \alpha_5 \Delta\, ln\,(M_t^e/V_t) + \alpha_6 \Delta\Phi_t + \alpha_7 \Delta\, ln\, A_t^e + \alpha_8 \Delta\, ln\, V_t + \epsilon_{2t}$$

$$\text{(15b)}\quad \alpha_1 < 0 \quad \alpha_2 > 0 \quad \gamma_1 < 0 \quad \gamma_2 > 0 \quad 0 < \alpha_4 < 1 \quad 0 < \alpha_5 < 1 \quad \alpha_6 < 0 \quad \alpha_7 < 0 \quad \alpha_8 = 1$$

The predictions on the parameters to be estimated are listed below each equation.[12] Note that I have attached a parameter α_8 to $ln\, V_t$ to test directly the hypothesis that it is unity.

Observe that the price equations contain a rather large number of expected normal variables which must be related to observable magnitudes. I assume that each expected normal variable can be expressed as a distributed lag of past values of the variable in question. This means that for any normal variable, for example, Q_t^e,

$$\Delta\, ln\, Q_t^e = \sum_{i=1}^{I} l_i \Delta\, ln\, Q_{t-i}$$

where the l_i are the relevant weights and where one would expect $\sum_i l_i = 1$.

[12] An alternative form for these equations would be to separate V_t from the other variables so that wage rates, raw material prices, etc., would appear in "nominal" rather than "real" terms. However, we will work below with the equations stated above for two reasons. First, the above approach more readily permits a direct test of the parameter α_8, which is an important aspect of the question of the impact of price expectations on prices. Secondly, the approach taken above reduces somewhat the severity of collinearity problems, especially between G_t and P_t. Note that the variables in these equations must be interpreted properly; H_{t-1} is a stock measured at the end of a period, P_{t-1} is an average over a period, etc. Also, the customary procedure is followed in converting the equations from continuous to discrete time.

Due to a limited sample size and to collinearity among the lagged values of the variable on which the distributed lag is formed, some restrictions on the weights must be imposed. In this study the Almon polynomial approach is employed to estimate the weights. This approach has the important advantage of allowing for separate distributed lags of different length and form for each of the expected normal variables in our price equations.

To apply the Almon approach, the degree of the polynomial and the length of the lag must be chosen for each normal variable. To maximize degrees of freedom, I used second-degree polynomials to approximate each set of weights. To select the appropriate lag lengths, I undertook a search process over various combinations of lag lengths and chose the combination which provided the best fit in terms of the standard error of estimate and the pattern of weights for the distributed lags.[13] This search process was undertaken for each of our price equations and for each sector and industry.

The only variable that was not estimated in this way is normal productivity. In this case, following other studies in the literature, it is assumed that technical change is labor saving which implies that $\alpha_7 = -\alpha_4$ and that normal productivity advance takes place at a constant rate equal to 3 percent per year. I then adjusted wage rates for normal productivity advances and used this adjusted wage rate variable in the empirical work. This also of course eliminates the need to estimate α_7.

B. *The Data*

The price equations were fitted to data for the manufacturing sector of the *U.S.* economy, using sector data (i.e., data for Total Manufacturing, Total Durables, and Total Nondurables), and data for selected two-digit industries, (i.e., Textile Mill Products, Paper and Allied Products, and Electrical Machinery).[14] Both sector and industry data were used to test the versatility of the model at different levels of aggregation.

The data are quarterly and cover the period from 1956-II through 1971-II for the sector equations, and from 1959-II through 1971-II for the industry equations. For the sector equations the starting point was selected to allow for lengthy distributed lags without becoming involved with the Korean War period. The starting point for the industry equations was dictated by the need for long lags and the lack of data on certain series (wage rates) before 1956. The ending point in both cases was chosen to avoid the disruption of the controls period.

For the sector equations, output prices are the relevant wholesale price indexes. For the industry equations, the price indexes are those compiled by Eckstein and Wyss. The latter are wholesale price indexes which were constructed to conform to the Standard Industrial Classification.

For factor input prices, I utilized the straight-time hourly wage rate as a measure for each sector and industry. For raw material prices, the spot price index for twenty-two industrial commodities was used for sector equations while the Eckstein-Wyss

[13]In general, I searched extensively for the combination of lag lengths that produced the lowest standard error of estimate for the equation as a whole. This is the procedure suggested by Peter Schmidt and Roger Waud. In several equations, however, there were several combinations of lag lengths that were close together and produced standard errors of estimate that were about the same. In these cases, I brought to bear judgment regarding the pattern of weights in the lag distribution as well as the standard error of estimate in selecting the best-fitting equation. Also note that no end-point constraints were used in estimating the lags.

[14]The industries were selected partly because of data availability and partly because of their production characteristics; Textile Mill Products tends to have fairly homogeneous products and should produce largely to stock, Electrical Machinery has products that are quite differentiated and should produce largely to order. Paper and Allied Products has a mixture of products and should contain elements of both kinds of production activity. See, Otto Eckstein and David Wyss for a useful description of these industries. These industries thus provide a test of the model under varying production conditions.

input price indexes were used in the industry equations.

Data for new orders, finished goods inventories, and unfilled orders were gathered from various issues of *Manufactor's Shipments, Inventories, and Orders*. These data were converted to real magnitudes by deflating by the relevant output price.

Finally, Moody's BAA bond rate was used as a measure of interest rates. As a measure of a general price level, the implicit price deflator for *GNP* was used for the sector equations, and the wholesale price index for total manufacturing was used in the industry equations. For "shift" variables, real national income was employed in the sector equations and real new orders for total manufacturing was employed in the industry equations.

C. *Results*

The results from estimating the price equations are presented in Tables 1 and 2. Table 1 presents the estimates of equation (14a) and Table 2 presents the estimates of equation (15a). It is useful to consider both tables at the same time in order to compare the results from estimating each of the price equations.

Consider first the impact of demand on prices. The results indicate a strong and

TABLE 1—EMPIRICAL RESULTS FROM ESTIMATING EQUATION (14a)[a]

	Equation	Parameter Estimates[b]							Summary Statistics[c]		
		$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_8$	R^2	S_e	DW
Sectors											
i	Total Manufacturing	.001	.007	.320	.865	.019	−.002	1.500	.772	.00265	2.12
		(.038)	(.019)	(.094)	(.342)	(.027)	(.012)	(.370)			
				[8]	[6]	[2]	[14]	[14]			
ii	Total Nondurables	−.046	.012	.670	.934	.137	−.025	1.725	.667	.00452	1.78
		(.040)	(.014)	(.319)	(.432)	(.066)	(.022)	(.723)			
				[10]	[6]	[4]	[10]	[8]			
iii	Total Durables	−.315	.115	.394	.952	.003	.006	1.505	.610	.01156	1.99
		(.121)	(.063)	(.140)	(.526)	.010	(.026)	(.592)			
				[6]	[2]	[2]	[6]	[6]			
Industries											
iv	Textile Mill Products	−.066	.010	.153	.662	−.061	−.017	.682	.740	.00608	1.91
		(.050)	(.021)	(.071)	(.421)	(.185)	(.013)	(.404)			
				[4]	[6]	[2]	[6]	[8]			
v	Paper and Allied Products	.042	.023	−.046	.849	.030	−.005	1.069	.632	.00543	1.97
		(.054)	(.033)	(.112)	(.480)	(.055)	(.023)	(.724)			
				[10]	[2]	[4]	[6]	[6]			
vi	Electrical Machinery	−.031	.065	−.005	.799	.027	−.006	1.203	.526	.00597	2.18
		(.040)	(.050)	(.068)	(.362)	(.094)	(.018)	(.523)			
				[6]	[2]	[2]	[6]	[8]			

[a]Seasonal dummies are not reported.

[b]The numbers without parentheses or brackets are parameter estimates. When a distributed lag is present, the number is a sum of distributed lag coefficients. Note that to interpret the sum as an estimate of the relevant theoretical parameter presumes that the expectational weights sum to unity. For example $\hat{\alpha}_3 = \alpha_3 \Sigma l_i$ where the l_i are the expectational weights that relate expected normal new orders to actual new orders; obviously $\hat{\alpha}_3 = \alpha_3$ only if $\Sigma l_i = 1$. It is impossible to estimate separately both α_3 and the l_i with polynomial distributed lags. The numbers in parentheses are standard errors. Each number in brackets is the length of the lag attached to the variable that is associated with the parameter that appears at the head of each column. When the lag is less than three, the lag is an unrestricted one; otherwise, the lag is a polynomial distributed lag. When no number appears in brackets, a lag beyond one period is not applicable.

[c]R^2 is the coefficient of determination, S_e is the standard error of estimate, and DW is the Durbin-Watson statistic.

TABLE 2—EMPIRICAL RESULTS FROM ESTIMATING EQUATION (15a)[a]

Equation	Parameter Estimates								Summary Statistics		
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_8$	R^2	S_e	DW
Sectors											
vii Total Manufacturing	.040	.049	−2.039	.110	.729	.015	−.057	2.83	.779	.00271	2.03
	(.040)	(.014)	(1.500)	(.228)	(.451)	(.009)	(.037)	(1.16)			
			[10]	[10]	[6]	[2]	[14]	[14]			
viii Total Nondurables	.017	.016	−.114	.267	.918	.196	−.009	.797	.706	.00442	1.78
	(.040)	(.014)	(1.582)	(.370)	(.874)	(.085)	(.039)	(1.399)			
			[10]	[10]	[6]	[4]	[10]	[10]			
ix Total Durables	−.170	.073	−2.784	.496	.495	−.089	−.061	2.869	.679	.01103	2.04
	(.105)	(.059)	(2.637)	(.501)	(1.521)	(.128)	(.058)	(1.922)			
			[6]	[6]	[4]	[2]	[6]	[6]			
Industries											
x Textile Mill Products	−.095	.043	−.017	.157	.964	−.152	−.019	.966	.784	.00579	1.98
	(.041)	(.014)	(.828)	(.090)	(.407)	(.172)	(.029)	(.854)			
			[6]	[6]	[6]	[2]	[6]	[6]			
xi Paper and Allied Products	.018	.037	−1.573	−.060	.786	.082	−.008	.692	.679	.00530	1.66
	(.053)	(.032)	(.635)	(.094)	(.501)	(.053)	(.052)	(.779)			
			[8]	[8]	[2]	[4]	[8]	[6]			
xii Electrical Machinery	.012	.044	−1.214	−.066	.831	−.070	−.046	2.522	.625	.00554	2.53
	(.036)	(.036)	(.612)	(.085)	(.342)	.090	(.029)	(.860)			
			[6]	[6]	[2]	[2]	[8]	[6]			

[a]See Table 1 fnn.

systematic influence of the expected normal level of demand on prices. In each equation, expected normal demand plays an important role, but the variables that capture the influence of this factor differ somewhat between the sector and the industry equations.

In the sector equations, the expected normal level of new orders, measured by a distributed lag of actual new orders, performs strongly. In each of the sector equations (see equations i–iii of Table 1) α_3 has the correct sign and is statistically significant (at least) at the .05 level. The result is of particular interest in that to my knowledge no other study has worked with distributed lags of new orders as a demand-oriented factor or has uncovered a demand-oriented factor that has consistently performed well across these sectors.

The use of expected normal new orders as a demand-oriented factor presumes that the price elasticity of sector demand is small. As a test of this assumption, consider the estimates of γ_1 for the sector equations (see equations vii–ix of Table 2). In each of these equations, while the estimate of γ_1 has the right sign, it is not significantly different from zero at the .05 level. This in effect justifies the use of expected normal new orders as an approximate summary measure of the impact of expected normal demand on prices.[15]

In the industry equations, although expected normal demand influences prices in each case, the appropriate specification of the variable differs from industry to industry. For Textile Mill Products, expected normal new orders again operates strongly and appears to be a reasonable summary measure of expected normal demand considerations (compare equation iv of Table 1

[15]Furthermore, observe that while the overall fit of equations vii–ix is somewhat better than equations i–iii, the confidence intervals of individual coefficients that are common to both sets of equations are in most cases considerably wider (see α_4, for example). This appears to be due to collinearity problems that arise when the distributed lag on new orders is replaced by distributed lags on the implicit price deflator and real income. Equations i–iii are thus preferable on these grounds as well.

with equation x of Table 2). For the other two-digit industries, however, this variable does not perform well in that the estimates of α_3 have the wrong sign and are insignificant. Nevertheless, expected normal demand is an important determinant of prices in these industries in that the relevant estimates of γ_1 are statistically significant and have the correct sign (see equations xi and xii of Table 2). This means that for these industries expected industry demand cannot be taken to be completely price inelastic, even approximately. A proper account of the influence of demand on prices requires that the determinants of expected industry demand be specified.[16] This conclusion should not be surprising. We should expect the price elasticity of expected industry demand to be small in sector data where the level of aggregation is very high, compared to two-digit industry data where the level of aggregation is comparatively low. In short, expected normal demand has an important impact on prices at the industry level, but the appropriate variables for capturing the effects of this factor is a question that will have to be decided on an industry by industry basis.

Finished goods inventories and unfilled orders are also demand-oriented forces that influence prices. The effects of these variables are embodied in the estimates of α_1 for inventories and α_2 for unfilled orders. The estimates of these parameters generally have the right sign, but they achieve statistical significance in only a few cases. Inventories and unfilled orders thus appear to have some influence on prices, but unlike expected normal demand they do not exhibit a strong and systematic influence on prices across sectors and industries.

As in conventional price equations, expected normal wage rates and to a lesser extent expected normal raw material prices consistently have a significant impact on prices.[17] In comparing the statistical performance of these cost-push forces with the demand-oriented forces that have been employed, however, the latter appear to perform on balance about as well as the former on the basis of statistical significance, correct signs, systematic effects across sectors and industries, etc.[18] Indeed, in some industries, for example, durable goods, the demand-oriented factors perform decidedly better. In this sense, contrary to popular views, our results suggest that demand-oriented factors are no less "important" than cost-push forces as determinants of prices.

Next, consider the issue of the *direct* impact of price expectations on prices. This is an area where very little empirical work has been done. Price expectations affect prices in two ways in our model. First, they do so through the influence of the estimated current normal average price level; this effect is captured essentially in the parameter α_8, which the model predicts to be unity. In testing for the presence of this factor, the results are quite favorable. The estimates of α_8 in our preferred equations—that is, in equations i–iv of Table 1 and xi and xii of Table 2 which are the relevant ones from the point of view of the demand-oriented

[16] I note, however, that my efforts in sorting out the appropriate shift variables for expected industry demand have not been successful. Equations x–xii of Table 2 report results with new orders for Total Manufacturing as the shift variable. I also experimented with real national income as a shift variable. Neither variable produced statistically significant coefficients for γ_2. More work is needed to isolate the appropriate shift variables in industries where the determinants of expected industry demand need to be specified.

[17] For several equations a small reduction in the standard error of estimate can be achieved by lengthening the lag on wage rates beyond the lags shown. This, however, produced implausibly high estimates of α_4. What should be stressed, however, is that extending the lag on wage rates would not change the conclusions regarding the impact of demand and price expectations on prices.

[18] The point estimates of the elasticities of the response of prices to changes in expected wage rates are somewhat higher than those for the demand-oriented factors, such as expected new orders. But, in comparing the quantitative impact of these factors on prices, it should be kept in mind that, for example, new orders are more volatile than wage rates. To illustrate, in Total Manufacturing over the sample period that was used for the regressions, an average of the absolute value of percentage changes in new orders is three times higher than that for wage rates.

forces—are generally statistically significant from zero. Although the point estimates of α_8 are somewhat high in several cases, they are in most cases within a standard error of unity, and in all cases they are not statistically different from unity at the .05 level of significance. These results provide support for the theoretical framework adopted in this paper whereby industries are assumed to be composed of firms who have some monopoly power and who set prices in accordance with the prices that they believe their competitors are currently charging.

A second way that price expectations affect prices in the model is through anticipated inflation in the normal industry price level; this effect is captured in the parameter α_6. In this case, the results are not so successful. In our preferred equations, while the estimates of α_6 generally have the predicted sign, they are statistically significant in only one equation. Hence, at least in the present study, there is little evidence that future price expectations have an important effect on price behavior.

The overall fit of the preferred price equations is quite good. The coefficients of determination in each case are quite high, especially for price equations in which the dependent variables are simple logarithmic differences of the basic price indexes. The Durbin-Watson statistics are generally within the acceptance region for the null hypothesis of no autocorrelation. However, these statistics may be biased due to the presence of lagged values of the dependent variable as an explanatory variable.

III. Summary

This paper has undertaken an analysis of the effects of demand and price expectations on price behavior. A model of price behavior has been formulated to pinpoint the demand-oriented factors that affect prices, and incorporate price expectations as a direct influence on price behavior. The model was fitted to data from the manufacturing sector of the *U.S.* economy. The major empirical results may be summarized as follows:

Expected normal demand has a strong and systematic influence on price behavior.

With sector data, the expected normal level of new orders effectively captures the influence of expected normal demand. But, with industry data, while expected normal new orders is an adequate measure of expected normal demand in some industries, the determinants of expected normal demand must be specified in others.

Inventories and unfilled orders have some effect on price formation, but their effects are felt sporadically across sectors and industries.

On balance, demand-oriented forces perform statistically about as well as cost-push forces as determinants of prices.

Expectations concerning the current normal industry price level have a significant and consistent impact on price behavior. This lends support to the notion that firms in setting prices take into account their estimates of the prices that are currently being set by their competitors.

Future price expectations appear to have little effect on current price formation.

REFERENCES

David Belsley, *Industry Production Behavior: The Stock-Order Distribution*, Amsterdam 1969.

Paul Earl, *Inflation and the Structure of Industrial Prices*, Lexington 1973.

O. Eckstein and R. Brinner, "The Inflation Process in the United States," study for the Joint Economic Comm., 92d Cong., 2d Session 1972.

——— **and D. Wyss,** "Industry Price Equations," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972.

Ray C. Fair, *A Model of Macroeconomic Activity*, Cambridge, Mass. 1974.

R. J. Gordon, "The Impact of Aggregate Demand on Prices," *Brookings Papers*, Washington 1975, *3*, 613–70.

H. A. John Green, *Aggregation in Economic Analysis*, Princeton 1964.

Lawrence R. Klein, *A Textbook of Econometrics*, 2d ed., Englewood Cliffs 1974.

D. Laidler and M. Parkin, "Inflation: A Survey," *Econ. J.*, Dec. 1975, *85*, 741–809.

L. J. Maccini, "An Aggregate Dynamic Model of Short-Run Price and Output Behavior," *Quart. J. Econ.*, May 1976, *90*, 177–96.

———, "An Empirical Model of Price and Output Behavior," *Econ. Inquiry*, Oct. 1977, *15*, 493–512.

W. D. Nordhaus, "Recent Developments in Price Dynamics," in Otto Eckstein, ed., *The Econometrics of Price Determination*, Washington 1972.

E. S. Phelps and S. G. Winter, "Optimal Price Policy under Atomistic Competition," in Edmund S. Phelps et al., eds., *The Microfoundations of Employment and Inflation Theory*, New York 1970.

H. Rose, "Effective Demand in the Long-Run," in James Mirrlees and Nicholas Stern, eds., *Models of Economic Growth*, New York 1974.

P. Schmidt and R. N. Waud, "The Almon Lag Technique and the Monetary Versus Fiscal Policy Debate," *J. Amer. Statist. Assn.*, Mar. 1973, *68*, 11–19.

A Model of Agenda Influence on Committee Decisions

By CHARLES R. PLOTT AND MICHAEL E. LEVINE*

Within a range of circumstances it appears to be possible to control a group's decision by controlling only the agenda. The boundaries of the range over which the agenda is such an overwhelmingly important parameter are not yet known, and the exact principles upon which the influence rests have not been identified. However, the research results reported below provide a first step in answering these questions.

Our approach to this problem originated in both practical and theoretical considerations. As a practical matter, we were involved in an important and complex committee decision. A large flying club in which we held membership was meeting to vote upon the size and composition of the aircraft fleet which would be available to the membership for flying. As members we had preferences about the fleet available to us and an opportunity to shape the agenda. Preliminary discussions and meetings had narrowed the range of possibilities greatly from hundreds of thousands of competing alternatives to a few hundred. Over these remaining possibilities, however, there were conflicting and strongly held opinions. The group was to meet once and decide the issue by majority vote.

Principles of economics and game theory suggest that the procedures and other institutional aspects of committee processes should be important in determining the outcome. Axiomatic social choice theory and voting theory also suggest the importance of these variables. Yet, models which characterize the subtle features of parliamentary procedures and the behavior they induce do not exist. Thus the practical problem was accompanied by an intriguing theoretical problem that presented us with the possibility of developing a mathematical theory of procedures and procedural influences on group decisions.

The meeting was held. The group used our agenda. The decision was the one we predicted.[1] With this apparent success, we then faced a perplexing problem of proof. Was the result a happy accident or was the decision a direct consequence of our efforts? In order to partially resolve this question, we turned to experimentation. If by using the methods we developed we were unable to influence groups involved in conflicts similar to the club meeting, then we would be willing to dismiss the club experience as an accident.

The experimental results below indicate that the club decision cannot be dismissed as accidental. The principles we outline for determining the agenda's influence are in need of improvement, but their fundamental importance within a range of circumstances is established. A more refined and accurate identification of the principles and the ranges over which they are operative awaits further research. Even as it stands our research has important implications for process evaluation and design (see the authors).

The paper is outlined as follows. In Section I, we outline a basic theory and a model. Section II includes our experimental design and Section III contains the results. The last section is a summary of conclusions.

*California Institute of Technology; and California Institute of Technology and University of Southern California Law Center, respectively. The research support provided by the National Science Foundation and the Henry Luce Foundation is gratefully acknowledged.

[1]The details of this meeting and a discussion of many of the problems of applications of the theory are reported in our referenced paper.

I. Theory and Model

First, we will develop a formal representation of an agenda. Then we will outline an intuitive theory about the nature of an agenda's influence, after which we will formally state a model.

A. *The Agenda*

The form of agenda we used in resolving the club problem can be represented abstractly as a series of partitions (into two sets) of the feasible set of alternatives. Each item on the agenda was designed to eliminate by majority vote some set of alternatives from further consideration. Our experimental agendas were similarly constructed.

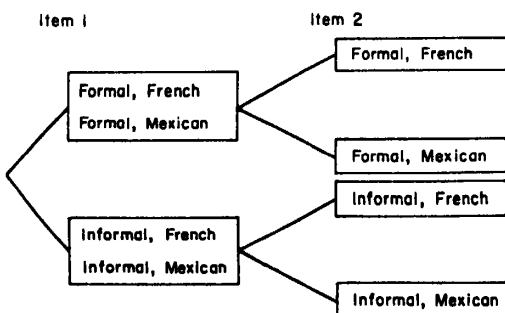
We used the following example to explain the agenda to subjects during some of our experiments. Suppose that we are deciding what kind of banquet to give. The agenda reads: Item 1. Shall the dress be formal or informal? Item 2. Shall the cuisine be French or Mexican? This agenda is modeled by Diagram 1. The vote is first on item 1 and then on item 2.

Each item on the agenda is designed to eliminate some of the alternatives which have survived the previous votes. This continues until a single alternative remains which is the choice of the group. For a fixed set of alternatives, the set of all agendas corresponds to the set of all such "trees," where each tree that can be formed from a given set of alternatives represents a different agenda. If, for example, the items above are reversed so the first vote is on cuisine and the second on attire, then the tree would be altered accordingly.[2]

DIAGRAM 1

B. *Basic Theory*

Our basic theory is simple. Where an agenda is fixed, it influences outcomes in two ways: first, it limits the information available to individual decision makers about the patterns of preference in the group. The primary means available for preference revelation is voting, and the content of each vote is specified by the form of the agenda. In some settings, other means of preference revelation such as verbal communication and/or straw votes can be ruled "out of order" by strict adherence to an agenda and therefore provide a limited means for information generation. And where there are many alternatives and many people, verbal communications may be of limited importance whether permitted or not. In addition, on-the-spot coordination of decisions among individuals through any type of binding agreement is nearly impossible in most meetings. This generally precludes expressly collusive behavior unless it is the result of a premeeting meeting and, even then, to be effective in planning strategy the coalition often needs to know both the patterns of preference among the group and the agenda to be used. Thus, each individual usually finds himself in a position of decision making under uncertainty. The preferences of others will have limited opportunity to influence his behavior.

Second, the agenda determines the set of strategies available to the individual. He always has the opportunity to choose among outcomes, but which outcomes he may choose among at any point is deter-

[2]It is always possible to represent a tree so that the corresponding agenda presents a set of choices the group will find acceptable or "natural"? We occasionally had to expend considerable effort on the wording of the agendas we used in experiments and suspect that some results cannot be reached using a natural appearing agenda. The agenda used during the club meeting is reproduced in the authors' paper.

mined by the agenda. The individual always must pick the particular strategy he prefers from among those available. The agenda determines what strategies are available. So, by reducing the influence of others' preferences and by determining the set of strategies available to him, the agenda effectively influences the voting pattern of each individual in the group. It thereby influences the choice made by the group.

C. *The Model*

The model is constructed to apply to a very broad range of circumstances as well as to our experimental setting. However, as will be explained below, certain very specific operational assumptions were made when applying the model in the experimental environment.

1. *Individual Voting Rules*

As indicated above an agenda item partitions the set of alternatives into two sets, one of which will be eliminated by vote. What decision rule will the individual use? We have postulated a universe limited to three rules.

Rule 1. *The sincere-voting hypothesis:* This hypothesis holds that an individual faced with two sets of alternatives will vote for the set which contains his most preferred alternative. If he is indifferent between the two best alternatives he then decides on the basis of a comparison between the second ranked alternatives. If he is indifferent between these two, then we define the rule to be ambiguous.[3]

Rule 2. *The avoid-the-worst hypothesis:* Here the individual votes to avoid the alternative he likes the least. When faced with a choice between two sets, he compares the least-preferred alternative in each set and votes against the set which contains the worst of these two. The case of ties is treated similarly to the above.

Rule 3. *The average value hypothesis:* This hypothesis holds that the individual treats the group choice as a lottery that will choose any alternative in a particular set with equal probability. The choice between two sets is like a choice between two lotteries (with uniform distribution over the outcomes) and he chooses (votes for) the one with the higher expected utility. The case of ties is treated as in rule 1 above.

Clearly, these three decision rules do not exhaust the set of imaginable decision rules. For example, the decision could also be affected by the variance of the payoff in a set, attitudes toward risk, past decisions made by the group, or subjective estimates of future decisions. If the model were to be refined further, this might be one of the places where it could be improved.

Our approach to the problem differs from that found in economics. We postulate the individual as a random variable over these decision rules. That is, we as experimenters do not know which rule he will use at a given point, but we are willing to speculate about the probability with which he will use a rule. In this "stochastic man" approach we are close to models which have had successful applications in marketing (see Frank Bass).

Some notation is needed.

Ω = the universal set of alternatives

$\mathfrak{a}$ = $(J_1, J_2, \ldots, J_m)$ is an *agenda* where J_k is a partition of each of the partitionable sets of J_{k-1} into two sets, and $J_0 = \Omega$

I = the set of all individuals

$u^i(x)$ is a von Neumann-Morgenstern utility function over Ω for $i \in I$

$\mathcal{S}(S, \bar{S})$ = $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9\}$ = the set of "states" in which an individual may find himself relative to two sets S and $\bar{S}$ of alternatives. These are defined as follows.

α_1 = All decision rules dictate a vote for S over $\bar{S}$; or one (or more) decision rule dictates a vote for S and the other two

[3]The hypothesis as first developed by Robin Farquharson continues in the lexicographic fashion. An ambiguity in his procedure can occur when the sets are of different sizes. This was called to our attention by Steven Matthews.

(or one) are ambiguous between S and $\bar{S}$.

α_2 = All decision rules dictate a vote for $\bar{S}$ over S; or one (or more) decision rule dictates a vote for $\bar{S}$ and the other two (or one) are ambiguous between $\bar{S}$ and S.

α_3 = One decision rule dictates a vote for S, another dictates a vote for $\bar{S}$, and the other is ambiguous between S and $\bar{S}$, or all three rules are ambiguous.

α_4 = Decision rule 1 dictates a vote for S and both rules 2 and 3 dictate a vote for $\bar{S}$.

α_5 = Decision rule 2 dictates a vote for S while both rules 1 and 3 dictate a vote for $\bar{S}$.

α_6 = Decision rule 3 dictates a vote for S while both rules 1 and 2 dictate a vote for $\bar{S}$.

α_7 = Both decision rules 1 and 2 dictate a vote for S while rule 3 dictates a vote for $\bar{S}$.

α_8 = Both decision rules 1 and 3 dictate a vote for S while rule 2 dictates a vote for $\bar{S}$.

α_9 = Both decision rules 2 and 3 dictate a vote for S while rule 1 dictates a vote for $\bar{S}$.

$P_i(S, \bar{S} | \alpha_k, \mathfrak{A})$ = the probability that individual i will vote for the set S over the set $\bar{S}$ given that they are imbedded at some stage in agenda $\mathfrak{A}$ and that he finds himself in the situation described by α_k.

AXIOM 1: *Independence from Environment: The probability distributions $P_i(S, \bar{S} \mid \alpha_k, \cdot)$ are parameterized only by α_k and for all $\hat{S}$, $\hat{\hat{S}}$, S', S'', $P_i(\hat{S}, \hat{\hat{S}} \mid \alpha_k) = P_i(S', S'' \mid \alpha_k)$.*

This means that the individual does not act strategically by anticipating upcoming votes; his probability of voting is not affected by previous votes; his probability is not affected by discussion at any stage of the meeting, set sizes, set labels, etc. It is as though he always uses one of the decision rules above, and he chooses from among them with fixed probabilities.

AXIOM 2: *Stochastically Identical Individuals:*

$$P_i(S, \bar{S} \mid \alpha_k) = P_j(S, \bar{S} \mid \alpha_k) \textit{ for all } i, j, S, \bar{S}, k$$

This axiom postulates a certain similarity among individuals. It says that the probability that any individual votes "yes" when he finds himself in any given situation is the same for anyone who finds himself in that same situation. In addition, this axiom declares that the universe of parameters on the probability distribution is exhausted by the situations enumerated above.

2. *The Strength of S against $\bar{S}$*

Suppose the voting rule is a majority rule and that in the agenda the set $\bar{S}$ has been pitted against the set S. What is the probability that S will win? This probability will be called the strength of S against $\bar{S}$ and can be calculated as follows.

$V(S, \bar{S}, \alpha_k)$ = the set of people who find themselves in situation α_k; $\alpha_k \in \mathcal{S}(S, \bar{S})$

N_k = the number of people in the set $V(S, \bar{S}, \alpha_k)$

n = the total number of people [note: $\sum_{k=1}^{9} N_k = n$]

W = $(z_1, \ldots, z_9)$: $z_i \in$ integers; and $0 \leq z_k \leq N_k$; and

$$n \geq \sum_{i=1}^{9} z_i \begin{cases} \geq \dfrac{n+1}{2} & \text{if } n \text{ is odd} \\ > \dfrac{n}{2} & \text{if } n \text{ is even} \end{cases}$$

$P(S, S)$ = the probability that the set S receives a majority vote over the set $\bar{S}$ in a contest between the two.

THEOREM:

$$(1)\quad P(S, \bar{S}) = \sum_{W} \prod_{k=1}^{9} \frac{N_k!}{(N_k - z_k)! z_k!} \cdot P(S, \bar{S} \mid \alpha_k)^{z_k} (1 - P(S, \bar{S} \mid \alpha_k))^{N_k - z_k}$$

That this is the appropriate probability can be seen by application of the binomial probability distribution, the independence assumptions and the appropriate area of summation. Notice that all we need to know to calculate this number is the number of people in each set $V(S, \bar{S}, \alpha_k)$ and the nine probability numbers represented by $P(S, \bar{S} \mid \alpha_k), k = 1, \ldots, 9$.

3. *Strength of an Agenda*

We turn now to the model of primary interest. What agenda is most likely to yield a given alternative x as the group's choice? We answer this question by calculating the strength of an agenda for an alternative x. We do this by first calculating the probability under a given agenda that x will be the group's choice. With that formula in hand, we can then survey all possible agendas (which is incidentally no simple problem) to find the one which maximizes the chance of getting x.

Consider the agenda $\mathfrak{A} = (J_1, \ldots, J_m)$. We assume there are m items. Each item J_k is a partitioning of each set in J_{k-1} into two sets. The original set $J_0 = \Omega$ is the set of all alternatives. Now since the items of the agenda are *partitions*, each element $x \in \Omega$ appears in one and only one set in any given item. Call this set $S(x, J_k)$ and the set which is pitted against it $\bar{S}(x, J_k)$.

Our previous formula (1) allows us to state the probability $P(S(x, J_k), \bar{S}(x, J_k))$ for any given x and any given J_k. From the independence axioms above we know immediately then that:

(a) $P(x \mid \mathfrak{A})$ = the probability that x is chosen by a group given that the agenda is $\mathfrak{A}$

(b) $P(x \mid \mathfrak{A}) = \Pi_{J_k \in \mathfrak{A}} P(S(x, J_k), S(x, J_k))$

This is the formula we were seeking at the beginning.

4. *Influencing the Group*

In order to apply the theory, we face four more problems. First, we must obtain preference estimates. The experiments explained below involved money payments. To simplify, we assumed that people were "risk neutral" so utility was linear in money payment. The second problem involves obtaining estimates of the nine numbers $P(S, \bar{S} \mid \alpha_k), k = 1, \ldots, 9$. The numbers we used were estimated from the pilot experiments and are provided in Section III.

The third problem involves the interesting mathematical problem of finding the optimum agenda. For each alternative we can compute the probability that it will win under any given agenda. Choice of an agenda then will be like the choice of a lottery so in general the "best" agenda would depend upon attitudes toward risk, etc. The objective function we use simply dictates finding the $\mathfrak{A}$ which maximizes $P(x \mid \mathfrak{A})$. The hard part occurs because of the *very* large number of potential $\mathfrak{A}$'s.

Fourth, we must be able to get the group to adopt and adhere to the agenda we have chosen. This involves devising an agenda which presents choices in an acceptable or "natural" way, preventing alternative motions from reaching the floor.

II. Experimental Procedures

We experimented by creating groups which had important features of the naturally occurring processes we wish to understand. We deduced these features from the club experience: 1) the group uses majority rule and a prearranged agenda which is followed closely; 2) there is little opportunity for premeeting meetings or predesigned coalitions to form prior to the meeting; 3) there is little or no uncertainty among the participants as to their attitudes toward the various candidate alternatives; 4) individuals are not indifferent among alternatives.

The first two conditions were easy to meet. Student subjects were recruited from Caltech, the University of Southern California, and the University of California-Los Angeles. An announcement was made in classes about the opportunity to participate in a "decision-making experiment." They were told that they would attend a meeting which would last approximately an hour, discuss some issue which had no political overtones, and that they would

have the *opportunity* to make "well over the hourly wage which any of them might be receiving." They were told that the experimenters were interested in certain logistical and technical problems about group decision processes; that there was no interest in psychological variables or personal variables; and that they would be subject to no harm or embarrassment.

Meetings took place in a classroom beginning at noon. As participants arrived they were assigned to seats in accord with a function which resulted from a random number table. When all participants were seated, they were asked to read the instructions which had been placed face down on their desks.

We adapted the theory of induced preference developed by Vernon Smith to take care of third and fourth conditions. The set of alternatives Ω was a subset of the letters of the alphabet. The task of the group was to use the appropriate procedures and choose one letter from this set. Each individual $i \in \{1, 2, \ldots, n\}$ was given a payoff function $u^i(x)$, $x \in \Omega$, which indicated the amount of money he would receive from the experimenter as a function of the alternative chosen by the group. He could not mention the amounts of money reflected by his payoff and no side payments, bribes, or threats were permitted. So, as long as an individual preferred more money to less, his preference relation over Ω is given by $xR_iy \Longleftrightarrow u^i(x) \geq u^i(y)$. In our case, the amounts involved seemed to us to induce well-defined preferences and nonindifference between alternatives. We assumed in addition that people were risk neutral.

The instructions were read by the experimenter, who did not know at the time which alternative the agenda was designed to produce. These are included in the Appendix. After reading the instructions the experimenter answered any questions, turned the meeting over to the chairman, and seated himself at the back of the room. He said nothing during the remainder of the experiment except when voting took place. He then stood up and recorded votes.

The chairman for Series 2, 3, and the final Series 4 was a Caltech senior majoring in physics. He was paid $4.00 per hour. He was given the instructions labeled "chairman's instructions" in the Appendix. He was not told the purposes of the experiment or that we had any expectations about which alternatives the group might choose. In the debriefing which occurred after the final experiment, it was evident that he did not know the purposes of the experiments and did not suspect that the agenda was a key variable.

The only person present during the experiment who was aware of which alternative was theoretically supposed to occur was the graduate research assistant, Steven Matthews. He was introduced along with the chairman, as a recording secretary. The only things he said during the meetings were functional to the general task of recording votes.

After the procedures had been fully discussed and the "test"[4] had been administered, the meeting began. The chairman took up the first item on the agenda and opened the floor for discussion. We asked him to encourage discussion on the first item. Participants tended to be a little hesitant to speak up ("What can I say about an *A*?"), but once discussion started, they often were moved to comment.

After the first item was voted upon, the group considered the next item on the agenda. On two or three occasions someone asked if items could be changed. This was not allowed. We suspect that certain types of straw votes are effectively changes in the agenda and may affect outcomes. Although we never prohibited a straw vote, we were prepared to rule one out of order if it was put in the form of a substitute agenda; for example, "If it comes down to box *A* versus box *B* later, how many will go for *A*?" We did allow one straw vote in this series and

[4]We found the test to be very useful. On several occasions during our pilot experiments we had reason to suspect that participants did not fully understand the agenda and/or motions. After we adopted this test, mistakes seldom occurred.

TABLE 1— ESTIMATES OF $P(S, \bar{S} \mid \alpha_i)$ MEASURED FROM SERIES 1-3

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9
.96	.04	.50	.38	.17	.61	.39	.83	.62

we think it did affect the outcome (see Table 2).

When the meeting was over, all subjects were paid in cash the amount dictated by their payoff sheet and the alternative chosen by the group.

III. Results

A total of four experimental series were conducted. The first three series, which are treated as pilots, served two functions. First, the procedures as reported here and the instructions (used in the fourth series) had been revised to take account of problems encountered in these first three series.

The second function of the pilot experiments was to provide data from which the probability parameters used in the model could be estimated. Both the numbers reported in Table 3 and the design of the Series 4 experiments were based on these estimates (see Table 1).

Series 4 consisted of four experimental sessions. The set of alternatives, Ω, contains five elements. The payoff schedules used in all four sessions are listed in Table 2. The majority-rule relation is also shown there. Alternative 1 beats all others in any binary contest and Alternative 5 is beaten (unanimously) by any of the others in a binary contest. The other three alternatives are involved in a cycle. For each of the first four alternatives, an agenda exists which would yield that alternative with a probability equal to one according to our model. We would have preferred to avoid the cycle, but we were unable to find a noncyclic example for which a probability one agenda could be constructed according to our model for each feasible[5] item, given the probabilities measured from Series 1.

[5]Alternative 5 is possible only with extremely low probabilities.

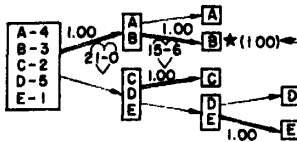
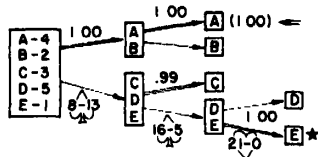
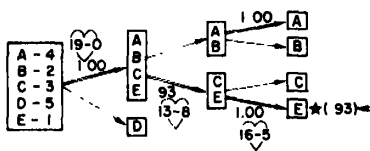
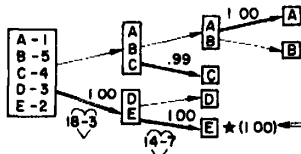
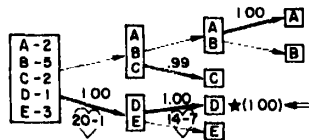
The results of these experiments are in Figure 1; Experiments 1, 3, and 4, which were designed to get Alternatives 3, 2, and 1, respectively, performed exactly as anticipated. Each resulted in the choice of alternatives for which the agenda was designed.

The agenda for Experiment 2 was designed for Alternative 4, but the group chose Alternative 1. This resulted because a straw vote revealed the fact that Alternative 5 (labeled D in this experiment) was least preferred by *all* individuals. Does this call into question the basic assumptions of our model? We think not. This straw vote, we claim, effectively changed the agenda to one on the figure labeled "Alternate Specification: Series 4 Experiment 2." For this

TABLE 2 SERIES 4 PAYOFFS IN DOLLARS

	Alternative				
Person	1	2	3	4	5
1	6.00	7.00	5.00	8.00	0.50
2	6.00	7.00	5.00	8.00	0.50
3	6.00	7.00	5.00	8.00	0.50
4	6.00	7.00	5.00	8.00	0.50
5	6.00	7.00	5.00	8.00	0.50
6	6.00	7.00	5.00	8.00	0.50
7	7.50	7.75	6.75	5.75	0.25
8	7.50	7.75	6.75	5.75	0.25
9	7.50	7.75	6.75	5.75	0.25
10	7.50	7.75	6.75	5.75	0.25
11	7.50	7.00	6.00	8.00	0.50
12	8.00	7.50	7.00	6.00	0.50
13	8.00	7.50	7.00	6.00	0.50
14	8.00	7.50	7.00	6.00	0.50
15	7.00	5.50	7.50	6.50	0.25
16	7.00	5.50	7.50	6.50	0.25
17	7.00	5.50	7.50	6.50	0.25
18	7.00	5.50	7.50	6.50	0.25
19	7.00	5.50	7.50	6.50	0.25
20	7.00	5.50	7.50	6.50	0.25
21	7.00	5.50	7.50	6.50	0.25

Majority Rule Relation: 1 ← 2 → 3 → 4 → 2 (cycle); 4 ← 5
1st: 1; cycle: 2, 3, 4; last: 5

SERIES 4 - Experiment 1

SERIES 4 - Experiment 2

Alternative Specifications
SERIES 4 - Experiment 2′

SERIES 4 - Experiment 3

SERIES 4 - Experiment 4

Key

(x) → means that x is the strength of the set over its complement as determined by the model

(x) ⟸ means that x is the strength of the agenda in getting this alternative. Each agenda was designed to get the alternative so marked

x-y (heart) means that at this stage the vote went "our way" by a vote of x to y.

x-y (spade) means that at this stage we "lost" by a vote of x to y

★ indicates the actual final choice by the group

FIGURE 1

alternate agenda the model predicts letter E, the one actually chosen with a .93 probability.

We now come to the most basic of questions. What are we prepared to say we have learned about our general theory and how can we easily summarize our beliefs? The following Bayesian argument is enlightening if we start from the two competing generalizations which existed before the research was initiated:

θ_0: The outcome of the process *does not* depend upon the agenda. That is, there exists a probability distribution $P(x)$ over outcomes $x \in \Omega$ which *is not* functionally dependent upon the agenda, although it may depend on other parameters.

θ_1: The outcome of the process *does* depend upon the agenda. That is, there exists a probability distribution $P(x \mid \mathfrak{a})$ over outcomes $x \in \Omega$ which *is* functionally dependent upon the agenda in addition to other parameters.

Cast in this framework the arguments in favor of θ_1 are very persuasive if we adopt the position of a critic who initially had low expectations about the truth of θ_1. Suppose, for example, we make the following assumptions where x is the observed sequence of outcomes:

i) The a priori probabilities are $P(\theta_0) = .8$ and $P(\theta_1) = .2$
ii) $P(x \mid \theta_0)$ is the maximum likelihood estimate .015625
iii) $P(x \mid \theta_1)$ is the prediction of the model

With these assumptions the a posteriori probability that θ_1 is true is .94. This critic is certainly impressed.

If our critic would not allow our explanation of Series 4 Experiment 2, then a repetition of the argument above would show that he has learned much less from Series 4. Our own priors which had resulted from observing pilot experiments were on the order of $P(\theta_1) = .9$ so without our Experiment 2 explanation, we learned very little from the experimental series. Since the cost of an additional experiment is about \$170 and any critic can study the pilot runs, we elected not to try to convince this critic until we found a setting within which we could learn something additional ourselves. We conclude that the agenda influences the outcome.

Even though our general theory may be right, the specific means of expressing or

TABLE 3: DISTRIBUTION OF OUTCOMES[c]

Series Experiment Item	Mean of Win Votes[a] μ	Standard Deviation σ	$\sqrt[3]{\mu_3}$	Number of Win Votes[a] x	$(x-\mu)/\sigma$	Probability of Direction Actually Taken	Probability of Final Outcome
1-1-1	11.57	1.67	-.48	12	.257	.90	
1-1-2	13.83	1.16	-.71	15	1.01	1.00	
1-1-3	10.84	1.71	-.65	11	.094	.79	.65
1-1-4	10.85	.95	-.47	11	.158	.93	
1-2-1	11.57	1.67	-.48	15	2.06	.90	
1-2-2	13.83	1.16	-.71	19	4.46	1.00	
1-2-3	10.84	1.71	-.65	12	.679	.79	.65
1-2-4	10.85	.95	-.47	13	2.27	.93	
1-3-1[b]	10.85	1.67	-.48	9	-1.54	.10	
1-3-2	10.54	1.15	-.41	11	.400	.83	.08
1-3-3	12.20	1.38	-.62	11	-.869	.97	
1-4-1	11.57	1.67	-.48	11	-.341	.90	
1-4-2	11.48	1.37	-.69	13	1.11	.93	.82
1-4-3	12.37	1.16	-.65	14	1.40	.99	
1-5-1	11.00	1.59	-.78	13	1.26	.83	
1-5-2	11.17	1.52	-.34	16	3.18	.87	
1-5-3	12.21	1.48	-.72	13	.534	.97	.58
1-5-4	10.40	1.05	-.40	10.5	.095	.82	
1-6-1	11.03	1.43	-.45	12	.679	.86	
1-6-2	10.58	1.46	-.64	11	.288	.78	
1-6-3	11.21	1.16	-.77	12	.682	.93	.61
1-6-4	12.62	1.42	-.88	13	.267	.98	
2-1-1	19.49	1.14	-.94	21	1.32	1.00	
2-1-2[b]	13.34	1.90	-1.11	10	-1.76	.07	.07
2-1-3	11.85	1.00	-.48	13	1.15	.92	
3-1-1	17.42	1.64	-1.10	20	1.58	1.00	
3-1-2(Item 5)[b]	11.85	1.00	-.48	10	-1.85	.08	.08
3-2-1[b]	19.26	1.21	-.98	8	-9.33	.00	
3-2-2	17.19	1.69	-1.13	16	.706	1.00	.00
4-1-1	18.34	1.44	-1.04	21	1.85	1.00	
4-1-2	13.65	1.00	-.64	15	1.35	1.00	1.00
4-2-1[b]	17.65	1.59	-1.09	8	-6.08	.00	
4-2-2[b]	16.73	1.78	-1.16	5	-6.59	.00	.00
4-2-3	19.95	.99	-.87	21	1.06	1.00	
4-2'-1	12.73	1.27	-.77	13	.213	.96	
4-2'-2	13.65	1.00	-.64	16	2.35	1.00	.96
4-3-1	17.65	1.59	-1.09	18	.220	1.00	
4-3-2(Item 3)[b]	13.65	1.00	-.64	14	.350	1.00	1.00
4-4-1	17.42	1.64	-1.10	20	1.58	1.00	
4-4-2	13.65	1.00	-.64	14	.350	1.00	1.00

[a]"Win" means that vote went in the direction indicated most probable by the model.
[b]These experiments did not result in the anticipated outcome.
[c]Data were pooled from all experiments except 1-3, 3-2, 4-2, 4-2'.

modeling it that we have developed is imperfect. First, the model made a probability one prediction which did not occur. Modifications to allow for straw votes may eliminate the problem. Secondly, we can, from Series 4, test the values of two parameters. The hypothesis that $P(S,\bar{S} \mid \alpha_1) = .96$, the number used in the model is accepted at the .01 level of significance. This is particularly interesting since it indicates that when individuals are in certain circumstances, our model of individual decisions is very good indeed. Psychological or other theoretical modifications are unnecessary. When all three rules cast compatible decisions, almost all behavior is explained. However, there were 32 votes cast from α_9 of which 27 were cast in the proper direction. According to the model these constituted 32 Bernoulli trials, each of which had a probability P of going in the proper direction. The hypothesis that $P(S,\bar{S} \mid \alpha_9) = .62$, the value used in the model, is rejected at the .01 level of significance. From this we know that our model could be improved by modifying the parameter values.

TABLE 4—CONSISTENT USE OF VOTING RULES
(Number of Individuals in Y^*)

Series and Experiment Number	Behavior Consistent with Rule 1	Behavior Consistent with Rule 2	Behavior Consistent with Rule 3	Behavior Consistent with None of the Rules	Total
1-1	1	1	6	2	10
1-2	4	1	1	4	10
1-3	0	0	0	1	1
1-4	0	0	6	0	6
1-5	1	0	1	2	4
1-6	1	0	3	2	6
Total	$\frac{7}{37}$ = .19	$\frac{2}{37}$ = .05	$\frac{17}{37}$ = .46	$\frac{11}{37}$ = .30	37

Note: Rule 1 = Sincere; Rule 2 = worst avoidance; Rule 3 = average value.
Y^* = the set of individuals for which the consistent use of any of the three voting rules would have been inconsistent with the consistent use of either of the other two voting rules.

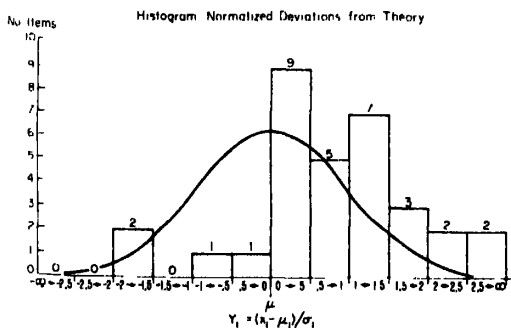
Table 3 provides a comparison between the actual vote and the predicted vote for each item of each experiment including all pilot series. The most significant thing about this table is the apparent conservatism in the model suggested by the very infrequent instances of the actual vote falling short of the expected vote (8 out of 40 cases). This conservatism shows up again on the histogram of Figure 2. If the theoretical distribution of votes for each item was normal, then the histogram should approach a normal distribution curve. But, for all items the theoretical distribution of votes was significantly skewed to the left (as shown by $\sqrt[3]{\mu_3}$ on the table). Since the histogram is strongly skewed to the right, the accuracy of the model is in even more doubt than the nonnormality of the histogram suggests. We suspect that this is a type of "bandwagon effect," but we have not tested for this.

FIGURE 2

Of particular interest to us were the patterns of individual decisions. Does an individual always use the same decision rule? Of the 261 individuals who participated in these experiments, only 37 were involved in a series of voting situations which would *necessarily*[6] reveal the individual's voting rule. Table 4 indicates 70 percent of these 37 subjects exhibited consistent behavior. The average value hypothesis was the most popular with about 46 percent of these subjects using it. The next largest group, 30 percent, used none of the rules consistently. The fact that so many individuals did not consistently use any of the rules suggests that some sort of probabilistic treatment of individual decision rules may always be necessary.

IV. Concluding Remarks

Our research incorporates several features not found (at least all in one place) in the economics and politics literature. First, our characterization of voting procedures is

[6]Any individual from among the 37 who consistently used any of the three rules would have exhibited behavior inconsistent with the use of either of the other two rules.

different from that found in the social choice and voting literature. With the exception of Farquharson, research in those areas focuses on processes in which alternatives are considered in a series of binary (two at a time) contests. The voting procedure we study involves voting between *sets* of issues. Our theory is decision theoretic in origin, but we depart from the traditional decision-theoretic mode of analysis by treating individuals as random variables over decision rules. Finally, our choice of an experimental methodology is certainly not typical of modes of analysis used by economists. Our posture is simple. If by using our ideas about the influence of the agenda, we are unable to influence the decisions of groups in a simple laboratory setting, then we cannot in good faith claim that our theory works in the more complicated "real world" case.

Experimental results indicate that within a range of circumstances the agenda can indeed be used to influence the outcome of a committee decision. Although the model we present needs improvement, the basic theory seems correct.

APPENDIX
INSTRUCTIONS USED IN SERIES 4

Chairman Instructions

You are employed to serve as chairman of several committee meetings. The time and location of these meetings are on the attached page. Each meeting will last about forty-five minutes. You should be at the designated location thirty minutes before the meeting starts and you should have familiarized yourself with the rules of order which are attached. For your participation you will be paid $5.00 per hour plus any necessary expenses, for example, parking, which you incur.

These meetings are part of a series of experiments designed to test theories about decision processes. Beyond this introductory remark, you will not be made aware of the purposes of the experiments until after the entire series has been completed. You should avoid talking with anyone about *any* aspects of the experiments, your employment, or about any possibly related theories. You should avoid circumstances in which you might inadvertently become informed. Do not try to guess the nature of the hypotheses or supply your own theories. After the final meeting you will receive a detailed explanation.

The first thing to do is check the dates and the times. Make sure you can be there. They are listed here as "Attachment No. 1." Attachment 2 is a copy of the instructions that members of the committee will receive. You should read these instructions now.

Here are some things that should be underlined:

1. People are free to say anything they wish which pertains to the motion on the floor. If discussions are "out of order," you can make that judgment. In particular, the following are not to be allowed:
   a) Statements which contain dollar or quantitative references;
   b) Straw votes on issues other than the current issue to be discussed and voted upon, as will be explicitly described on the agenda; and
   c) Threats or dealings between committee members to be carried through, during, or after the experiment is over.
2. Majority rule means a majority of those present. A vote passes if it receives 11 or more votes. If an item on the agenda fails both votes, you call for more discussion. After discussion another vote is taken. If neither passes you move to the next item on the agenda. An ambiguity after all items on the agenda are covered, can be resolved by a motion from the floor.

Parliamentary Rules for Chairman

Read the appropriate portions at the appropriate times.

Recognition Rule: Raise your hand to be recognized by the chair.

Voting Rule: The basic voting rule is simple majority rule. An issue passes if it passes by a majority of those voting.

Rule to Break Ties (read this if neces-

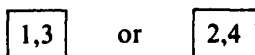
sary): If a tie vote occurs, discussion of the motion is again opened. After debate a second vote is taken. If a tie occurs again, debate is opened again and a vote is taken. If a tie occurs again, the committee moves to consider the next issue. Any ambiguity at the end of the last item can be removed by a motion from the floor.

Rule to End Debate: If someone wishes to end the debate on an item they simply move to end debate. If there is no objection to ending debate the item is voted upon.

(Read if necessary): If there is objection to ending debate, the motion to end debate will be recognized by the chair. A vote on the motion to end debate will be taken. If it passes by 2/3 majority of those voting the debate ends. If the motion to end debate fails, debate on the main motion continues.

AGENDA: The agenda committee has adopted the agenda which is before you. Notice that each item on the agenda is designed to restrict the number of programs which may receive further consideration. Example: Choice of banquet

Alternative	Type of Food	Dress
1	Mexican	Formal
2	Mexican	Informal
3	French	Formal
4	French	Informal

Item 1. Shall we have a formal dress banquet or not? Notice that an answer to this question will restrict further deliberation to either

Item 2. What type of food? Notice that an answer to this question is now all that we need to decide upon a specific alternative, as shown in Diagram 2.

Instructions for Committee Members

1. We would like for you to participate in a committee process experiment. The purpose of the experiment is to help us understand certain technical aspects of the generally complex ways in which committees operate. Support for this research was supplied by the National Science Foundation and the Henry Luce Foundation.

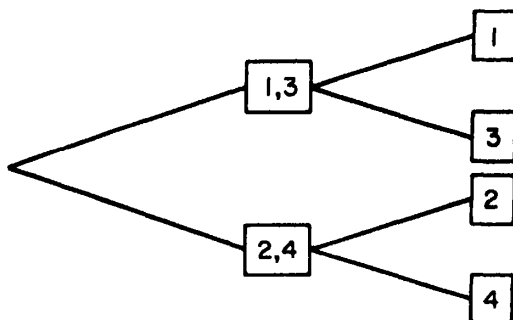



DIAGRAM 2

2. All you have to do is attend a committee meeting and for this participation you will be paid. The purpose of the meeting is to choose by majority rule a letter from the set of letters [*A*, *B*, *C*, *D*, *E*]. Only one of the five letters will be chosen and the payment you receive for participation depends entirely upon which one it is. For example, on the table on page 3, the amount listed beside the letter *A* is the amount you will receive if it is chosen by the committee; the amount beside *B* is the payment you will receive if it is the majority decision, etc.

Different individuals will receive different payoffs depending upon which letter the committee chooses. The letter which would result in the highest payment to you may not result in the highest payment to someone else. You should decide after deliberation how you wish the committee to vote and make whatever efforts you might want to get the vote to go that way. However, in general, we as experimenters are not concerned with whether or how you participate in the committee's effort to select a letter.

We want the meeting to proceed in an orderly fashion so we have provided a few parliamentary procedures which must be followed. These will be explained by the chairman. We also want to make sure that you understand the consequences of your

votes and any resulting committee decision. For this purpose we ask you to answer the question on page 4 after the chairman has reviewed the rules and the agenda.

3. Here are some incidentals:

a) The basic procedure will be simple majority rule. We will also follow the agenda prepared by an agenda committee. This agenda is outlined on page 3 and should be studied carefully. It will also be covered by the chairman.

b) You will from time to time be voting. We have appointed a recording secretary to record all votes. This can take some time so we ask you to hold your hands high until all votes are recorded.

c) You will be paid in cash immediately after the meeting. You may not reveal any *quantitative* information about your payment. If you wish you can say that one yields more than another, but you may not say how *much* more. The amounts may differ among committee members and *only you* are to know anything about how *much* you may receive.

d) Before or during the meeting please do not discuss with other committee members any activity to take place after the meeting which may involve you jointly. Under no circumstances may you make threats or "deals" to split your payment from the meeting with another committee member.

4. Are there any questions?

SERIES 4: EXPERIMENTS 1 AND 2

Individual Payment and Agenda Section of Individual Instructions. Committee Member ______.

Letter	Payment to you
A	
B	
C	
D	
E	

AGENDA are shown in Diagram 3.

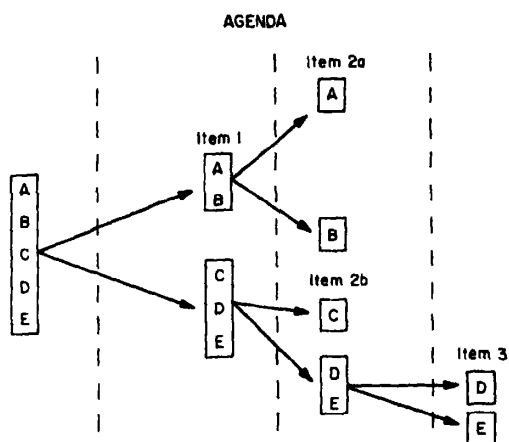



DIAGRAM 3

Item 1. Do we want to consider further only the letters *A* and *B*, or only the letters *C*, *D*, and *E*? (Check your vote.)

____I am in favor of considering further only the letters *A* and *B*.

____I am in favor of considering further only the letters *C*, *D*, and *E*.

Item 2a. (If the letters *A*, *B* are chosen at Item 1, then this item is applicable—if not, then go to 2b.) Which do we want, *A* or *B*?

____I am in favor of *A*.

____I am in favor of *B*.

Item 2b. (If the letters *C*, *D*, and *E* are chosen at Item 1, then this item is applicable -otherwise go to 2a.) Do we want to consider further only the letters *D* and *E*, or do we want to stop with *C*?

____I am in favor of *C*.

____I am in favor of considering further only the letters *D* and *E*.

Item 3. Do we want *D* or *E*?

____I am in favor of *D*.

____I am in favor of *E*.

Agenda Test Section of Individual Instructions

1. Suppose the top box at Item 1, the one that contains the letters *A* and *B*, received a majority of the votes, then the next item to be considered on the agenda is ____, and it consists of a vote between the letter(s) ____ and the letter(s) ____.

2. Suppose at Item 1 the box of letters that contains the letters *C*, *D*, and *E* is chosen by a majority. Then the next item to to be considered on the agenda is ____, and it consists of a vote between the letter(s) ____ and the letter(s) ____.

3. If the box of letters that contains *A* and *B* received a majority vote at Item 1, would there be a vote at Item 3? Answer Yes or No: ____. If it happened that the box of letters containing *C*, *D*, and *E* received a majority of votes at Item 1, and a vote was not needed at Item 3, then the box containing the letter ____ must have received the majority of votes and thus would be the committee's final choice.

4. If at each item the lower arrow was followed by the majority of votes, then the committee will have made ____ the final choice and you will receive the amount ____ as your payoff.

5. How much will you receive if the committee's final choice is: *D*?____ *B*?____ *C*?____ *A*?____

SERIES 4: EXPERIMENTS 3 AND 4

Individual Payment and Agenda Section of Individual Instructions. Committee Member No. ______.

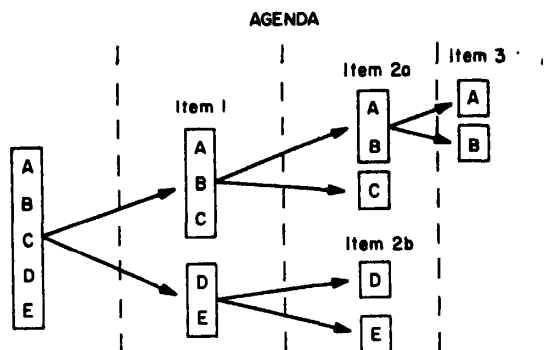
Letter	Payment to You
A	
B	
C	
D	
E	

AGENDA are shown in Diagram 4.

Item 1. Do we want to consider further only the letters *A*, *B*, and *C*, or only the letters *D* and *E*? (Check your vote.)

____ I am in favor of considering further only the letters *A*, *B*, and *C*.

____ I am in favor of considering further only the letters *D* and *E*.

Item 2a. (If the letters *A*, *B*, and *C* are chosen at Item 1 then this item is applicable —if not then go to 2b.) Do we want to consider further only the letters *A* and *B*, or do we want to stop with *C*?

____ I am in favor of considering further only the letters *A* and *B*.

____ I am in favor of *C*.

Item 2b. (If the letters *D* and *E* are chosen at Item 1, then this item is applicable—otherwise go to 2a.) Which do we want, *D* or *E*?

____ I am in favor of *D*.

____ I am in favor of *E*.

Item 3. Do we want *A* or *B*?

____ I am in favor of *A*.

____ I am in favor of *B*.

DIAGRAM 4

Agenda Test Section of Individual Instruction

1. Suppose the box at Item 1, the one that contains the letters *A*, *B*, and *C*, received a majority of the votes. Then, the next item to be considered on the agenda is ____, and it consists of a vote between the letter(s) ____ and the letter(s) ____.

2. Suppose at Item 1 the box of letters that contains the letters *D* and *E* is chosen by a majority. Then the next item to be considered on the agenda is ____, and it consists of a vote between the letter(s) ____ and the letter(s) ____.

3. If the box of letters that contains *D* and *E* received a majority vote at Item 1, would there be a vote at Item 3? Answer Yes or No ____. If it happened that the box of letters containing *A*, *B*, and *C* received a majority of votes at Item 1, and a vote was not needed at Item 3, then the box contain-

ing the letter ____ must have received the majority of votes and thus would be the committee's final choice.

4. If at each item the lower arrow was followed by the majority of votes, then the committee will have made ____ the final choice and you will receive the amount ____ as your payoff.

5. How much will you receive if the committee's final choice is:

D? ____ *B*? ____ *C*? ____ *E*? ____

REFERENCES

Kenneth J. Arrow, *Social Choice and Individual Values*, 2d ed., New York 1966.

F. M. Bass, "The Theory of Stochastic Preference and Brand Switching," *J. Marketing Res.*, Feb. 1970, *11*, 1–20.

Robin Farquharson, *Theory of Voting*, New Haven 1969.

Peter C. Fishburn, *The Theory of Social Choice*, Princeton 1973.

M. E. Levine and C. R. Plott, "Agenda Influence and Its Implications," *Virginia Law Rev.*, May 1977, *63*, 561–604.

C. R. Plott, "Axiomatic Social Choice Theory: An Overview and Interpretation," *Amer. J. Polit. Sci.*, Aug. 1976, *20*, 511–96.

Amartya K. Sen, *Collective Choice and Social Welfare*, San Francisco 1970.

V. Smith, "Notes on Some Literature in Experimental Economics," soc. sci. work. paper no. 21, California Inst. Technology, Jan. 1973.

Conglomerate Mergers, Default Risk, and Homemade Mutual Funds

By CORRY AZZI*

There are two common conjectures about the causes of conglomerate mergers: corporate shareholders prefer distributions of returns that can be acquired only from portfolios of the stocks and bonds of the conglomerate; and merger reduces the risk of investments. These conjectures have not been widely accepted as satisfactory explanations of conglomeration. A counter conjecture appears in the literature. If it were proven, it properly could be called the homemade mutual fund theorem. This theorem would state that any return distribution from a portfolio containing the debt and equity of a conglomerate could have been acquired through a portfolio containing some combination of the securities of the distinct corporations. Vernon Smith conjectures that "... in the absence of economies in the joint development of two or more investment activities or economies of scale in financing, corporate mergers (particularly 'conglomerates') would be distinctly undesirable. Greater flexibility and choice are provided by letting investors put together their own private 'merger' combinations on personal account by their choice of portfolios" (p. 461).

Smith claims that the merger of firms in distinct risk classes restricts an investor's opportunity set. But the effect of merger depends not only on the opportunities provided by the conglomerate's securities. It also depends on other securities or credit transactions that may be available to replicate or undo the merger. If opportunity sets are not changed, investors would be indifferent to the merger, and the conglomerate's value would be the sum of the values of the firms before the merger.

If the securities market corresponded to an Arrow-Debreu market, then the opportunity set would be unaffected. Likewise even in the presence of default risk, if investors can borrow with limited liability, pledging as sole collateral only the securities purchased in part with the loan, then mergers can be replicated or undone on personal account. But if the merging firms are not in the same risk class, then they and the conglomerate that they would comprise are all in distinct risk classes. Limited liability personal loans would be sufficient to ensure that investors could create homemade conglomerates or undo mergers, if each of the two firms and the conglomerate were not the only members of their respective risk classes.

In any event, whether the merging firms are or are not in distinct risk classes, the conditions sufficient to ensure that an investor's opportunity set would be unaffected are straightforward generalizations of conditions on security or credit markets sufficient for the homemade leverage theorem.[1] But such conditions, which are especially strong if default risk is allowed, would be relevant only if the conglomerate's securities make available opportunities that are somehow distinct from the opportunities that would otherwise have been available from the securities of the firms that comprise it.

Assuming no default risk on the bonds of both corporations that are party to a merger, I will prove the homemade mutual fund theorem and show that Smith's conjecture is essentially correct. If firms in distinct risk classes merge, investors cannot undo the effects of the merger through adjustment only in their holdings of the conglomerate's se-

*Assistant professor of economics, Lawrence University. I owe a considerable debt to Vernon Smith for his work on corporate financial theory. Steven Arnold, David Baron, and James Cox made very helpful comments.

[1]See David Baron, Kåre Hagen, Franco Modigliani and Merton Miller, and Joseph Stiglitz for detailed discussions of homemade leverage.

curities. But investors could always replicate the effects of mergers through homemade mutual funds. Thus merger could not enlarge but may restrict an investor's opportunity set, depending on the structure of the securities market. More importantly, the homemade mutual fund theorem does not generalize to include default risk. The conglomerate's securities would yield opportunities not available through any diversified portfolio of the distinct corporations' securities.[2] Merger would not restrict but could change an investor's opportunity set, again depending on the other securities or credit transactions that may be available to the investor.

Some authors have asserted that merger makes stocks and bonds less risky and thereby improves access to capital markets.[3] Implicit in this argument is the assertion that merger cannot lower but may raise firms' market values. I will show that this assertion is a nonsequitur. Even if firms subject to default risk merge and investors' opportunity sets are changed, the merger does not provide any opportunity that is necessarily less risky than the opportunities that would have been available through a diversified portfolio of the distinct corporations' securities prior to merger. Since in the presence of default risk, return distributions may be changed but could not necessarily be made less risky, investors' preferences may have to be invoked to explain merger activity.

I. Default Risk and Return Distributions

In this section the effects of merger on return distributions will be analyzed. The cash flow of either of the distinct corporations $j = 1, 2$ is $S_j(1 + \theta_j)$, where S_j is the amount of the corporation's capital expenditure, and θ_j is a random variable. The corporation has raised S_j in funds by sales of Y_j units of stock, each with a market value of one dollar, and bonds with a market value of Z_j and a redemption value of $Z_j(1 + r_j)$. Because of limited corporate liability, the random variable is not a rate of return on capital expenditures for $\theta_j < -1$. The corporation earns a rate of return of -1 if $\theta_j < -1$, but defaults on obligations incurred through production and trade. It has a negative cash flow. For ease of exposition, the analysis in this section will be done with the assumption that $pr\{\theta_j < -1\} = 0$. This does not alter the essential result that the homemade mutual fund theorem requires no default risk on bonds. The effects of admitting negative cash flows will be indicated summarily.

Since the desired explanation for conglomerate mergers is not to depend on scale economies or market power, assume that the probability of observing any (θ_1, θ_2) is not changed by merger. Furthermore, since only the financial consequences of a merger are to be assessed, the conglomerate is assumed to be created by a simple exchange of assets. Stock premiums and the opportunity to pay-in capital to the conglomerate are excluded without loss of generality. Unless merger can be shown to somehow change investors' opportunity sets, the willingness of some investors to pay stock premiums to others or to pay-in capital only in the event of conglomeration could not be explained.

Since a unit of stock from either of the distinct corporations is defined as one dollar's worth of stock, a unit of either corporation's stock would be exchanged for

[2]Even if investors' opportunity sets are changed by merger, the sets are not disjoint. If the return distributions on the investors' equilibrium portfolios are contained in the intersection of the sets, then investors would be indifferent to the merger. I am indebted to Baron for pointing out to me that the property sufficient to ensure that an investor's optimum is contained in the intersection of the sets is stronger than separation. As will be demonstrated, the intersection of the sets contains only the return distributions that an investor could acquire if he held the same percentage of all securities issued by the distinct firms. The percentage would be the same if an investor's preferences are representable by a quadratic utility function. See David Cass and Stiglitz for a detailed discussion of separation.

[3]The literature on conglomeration does not appear to be informative about the purely financial effects of merger precisely because diversified portfolios have not been considered as alternatives to merger. See, for example, Morris Adelman and Michael Gort.

one unit of the conglomerate's stock. Likewise, each dollar of redeemable value in the bonds of either distinct corporation yields one dollar of redeemable value in the conglomerate's bonds. Whether the original bonds are redeemed or remain outstanding is irrelevant, but the merger does not affect total obligations to bondholders. Thus

$$(1) \qquad Z(1 + r) = \sum_j Z_j(1 + r_j)$$
$$Y = \sum_j Y_j$$

where Y is the number of shares outstanding and Z is the market value of debt that yields interest rate r. Since the conglomerate's bonds and stocks may be freely traded, Z and the market value of Y are not prescribed by the terms of a merger agreement.

In order to assure that the purely financial consequences of a merger are not confused with the consequences of changes in corporate policy that may arise after a merger is consummated, assume that the scale of the conglomerate's activities S would be such that $S = S_1 + S_2$, where S_1 continues to yield θ_1 and S_2 continues to yield θ_2. Although the market value of the conglomerate's securities need not be the sum of the market values of the distinct corporations' securities, the scale of the conglomerate's activities are assumed to be determined by the market values of the distinct firms. This assumption is not restrictive. If for all possible S_1 and S_2 investors' opportunity sets are not changed by merger, investors would have no reason for promoting a merger and then changing corporate policies as compared to changing policies of the distinct corporations and then suitably diversifying portfolios.

To put the model in its proper context, assume that an investor is confronted with the possibility of two worlds. In the first, the two firms are distinct; the investor can hold a diversified portfolio of bonds with market values z_1 and z_2, redemption values $z_1(1 + r_1)$ and $z_2(1 + r_2)$, and stocks in amounts y_1 and y_2. In the second, the firms have merged, and the investor can hold bonds with a market value of z and a redemption value of $z(1 + r)$ along with an amount of stocks y. Stock prices and prices of a dollar of redeemable bond value may differ significantly between the two worlds, if return distributions on portfolios available in one differ from those in the other. But in order to determine if available return distributions can differ, return distributions on portfolios (z, y) have to be compared to return distributions on (z_1, z_2, y_1, y_2).

If either of the distinct corporation's rate of return on capital expenditures is less than the default rate of return, the minimum rate it must earn to pay its debt obligations, bondholders share all the corporation's returns and shareholders receive nothing. The default rate of either corporation is

$$(2) \qquad \theta_j^* = (r_jZ_j - Y_j)/S_j \qquad \text{for } j = 1, 2$$

Any investor holds z_j and receives z_j/Z_j of total returns available to a defaulting corporation's bondholders, and therefore receives

$$(3) \qquad R_j = S_j(1 + \theta_j)z_j/Z_j \qquad \text{for } j = 1, 2$$

from the holdings of the corporation's bonds.

If a corporation is not in default, any investor receives the redemption value of his bonds plus a share of income on equity. Income on equity is the corporation's total returns less its debt obligations, and an investor's share of that is y_j/Y_j. The investor receives

$$(4) \qquad R_j^* = z_j(1 + r_j) + [(S_j(1 + \theta_j) - Z_j(1 + r_j))y_j/Y_j] \qquad \text{for } j = 1, 2$$

An investor's income is $W + N$; N is assumed to be nonrandom and to consist of wage income and return on riskless assets. Statements (2)-(4) imply that

$$(5)\qquad W = \begin{cases} \sum_j R_j & \text{for } \theta_j < \theta_j^* \\ & j = 1,2 \\ R_i + R_k^* & \text{for } \theta_i < \theta_i^* \\ & \theta_k \geqq \theta_k^* \\ & i,k = 1,2 \\ & \text{and } i \neq k \\ \sum_j R_j^* & \text{for } \theta_j \geqq \theta_j^* \\ & j = 1,2 \end{cases}$$

Returns W are to be compared to returns an investor could receive from the conglomerate, which has a cash flow per dollar of capital expenditure of $(1 + \theta)$,

$$(6)\qquad \theta = \sum_j \alpha_j \theta_j$$

and α_j is the proportion of the conglomerate's capital expenditures S that yield θ_j.

$$(7)\qquad \alpha_j = S_j/S \qquad \text{for } j = 1,2$$

where $\sum_j \alpha_j = 1$. The conglomerate has just enough income to cover its debt obligations if

$$(8)\qquad \sum_j S_j(1 + \theta_j) = \sum_j Z_j(1 + r_j)$$

Statements (2) and (6)–(8) are used to find that the default rate of return for the merged corporation is

$$(9)\qquad \theta^* = \sum_j \alpha_j \theta_j^*$$

If the conglomerate is in default, an investor receives a share of returns that is proportional to his holdings z of the corporation's debt obligations. An investor gets

$$(10)\qquad R = \sum_j S_j(1 + \theta_j)z/Z$$

If the corporation is not in default, the investor gets the redemption value of bonds plus income on equity that is proportional to his holdings y. The investor receives

$$(11)\qquad R^* = z(1 + r) + \sum_j (S_j(1 + \theta_j) - Z_j(1 + r_j))y/Y$$

Income is the sum of portfolio income W', and other income N. Statements (6) and (9)–(11) imply that

$$(12)\qquad W' = \begin{cases} R & \text{for } \theta_j < \theta_j^* \qquad j = 1,2 \\ R & \text{for } \theta_i < \theta_i^* \text{ and } \theta_k \geqq \theta_k^* \\ & \text{which imply } \theta < \theta^* \\ & i,k = 1,2 \qquad i \neq k \\ R^* & \text{for } \theta_i < \theta_i^* \text{ and } \theta_k \geqq \theta_k^* \\ & \text{which imply } \theta \geqq \theta^* \\ & i,k = 1,2 \qquad i \neq k \\ R^* & \text{for } \theta_j \geqq \theta_j^* \qquad j = 1,2 \end{cases}$$

Propositions about the effects of merger on portfolio returns can now be proved. The first theorem identifies portfolios (z, y) and (z_1, z_2, y_1, y_2) that would yield the same distributions of returns. The theorem will be proved using two hypotheses: **H.1** $pr\{\theta_j < -1\} = 0$; **H.2** for each range of random variables listed in (12), there exists at least three ordered pairs of values of the random variables that have nonzero probabilities and are not linearly related.[4] For example, consider the values of the random variables taken on an interval in two dimensions such that $-1 \leq \theta_1 < \theta_1^*$ and $-1 \leq \theta_2 < \theta_2^*$. The hypothesis implies that at least three ordered pairs of values on the interval will have positive probabilities and will not be linearly related. The hypothesis establishes that the two firms are in distinct risk classes, and by statement (6), the definition of θ, the two firms and the conglomerate that they would comprise are all in distinct risk classes.

THEOREM 1: *Given* **H.1** *and* **H.2**, *portfolios* (z_1, z_2, y_1, y_2) *and* (z, y) *yield the*

[4] If the random variables are assumed to be continuously distributed, then the hypothesis can be suitably modified to refer to intervals on which a joint density function assigns probability.

same returns for all θ_i if, and only if $y/Y = z/Z = y_j/Y = z_j/Z, j = 1, 2$.

The proof of the theorem is found in the Appendix, but the logic is intuitive and follows from inspection of (5) and (12), which show that R_j, R_j^*, R and R^* are linear functions of θ_j; therefore **H.2** is used to argue that $W = W'$ for θ_j if and only if all coefficients on the random variables are the same.

A natural interpretation of Theorem 1 is that it applies to investors who take long positions in stocks and bonds; however the theorem easily generalizes to include margin risk and short sales under the assumption of an investor's personal liability up to the limit of income from all sources. Short sales can be interpreted as $y_i < 0$ under the supposition of personal liability for the cash flow generated by the stocks. Margin risk imposes a fixed interest rate obligation on borrowed funds and that obligation is a component of N, nonrandom income. The only significant change in the theorem that could be caused by margin risk is a lower bound on θ_i below which income from all sources less personal liabilities is zero. Short sales imply a finite upper bound on θ_i above which net income is zero. However, the assumption of linear independence of random variables within the bounds imposed by margin risk and short sales would leave the theorem essentially unaffected, if investors are personally liable.

Two corollaries follow from Theorem 1 and, depending on the nature of security markets, identify costs and benefits from conglomeration rather than portfolio diversification. Consider any investor who holds bonds with a market value of z_j and a redemption value of $z_j(1 + r_j)$ and stock of an amount of y_j. As a consequence of merger that investor receives a claim on income from y shares of the conglomerate's stock and bonds with a market value of z and a redemption value of $z(1 + r)$ such that

$$(13)\quad y = \sum_j y_j; \quad z(1 + r) = \sum_j z_j(1 + r_j)$$

The issue in the analysis of conglomerate mergers is what does merger provide for an investor that cannot be provided through portfolio diversification. An investor's optimal diversified portfolio, z_j^0 and y_j^0, generates a portfolio of the conglomerate's securities of bonds with a redemption value of $z^0(1 + r) = \sum_j z_j^0(1 + r_j)$ and of shares $y^0 = \sum_j y_j^0$. (Note the slight abuse of notation.) Could the investor have held a diversified portfolio $\hat{z}_j$ and $\hat{y}_j$ that yields the same returns for all θ_j as z^0 and y^0 of the conglomerate's securities? Since a unit of stock of either distinct corporation sells for one dollar, if a z_j and y_j exist such that $\sum_j(\hat{y}_j + \hat{z}_j) \leqq \sum_j(y_j^0 + z_j^0)$, then the investor has already revealed a preference for a diversified portfolio of z_j^0 and y_j^0 to z^0 and y^0. More importantly, if for each possible z^0 and y^0, there exist $\hat{z}_j$ and $\hat{y}_j$, then no investor can acquire a portfolio of the conglomerate's securities that is preferred to the diversified portfolio z_j^0 and y_j^0.

In the presence of default risk, there does not exist a $\hat{z}_j$ and $\hat{y}_j$ for each possible z^0 and y^0. This conclusion is stated as

COROLLARY 1: *Given* **H.1** *and* **H.2**, *there exists a portfolio containing $\hat{z}_j$ and $\hat{y}_j$ of securities in the separate corporations satisfying*

$$\sum_j (\hat{y}_j + \hat{z}_j) \leqq \sum_j (y_j^0 + z_j^0)$$

and such that returns on the portfolio would be the same for all θ_j as returns on a portfolio containing z^0 and y^0 of the merged corporation's securities if $y^0/Y = z_j^0/Z_j$, $j = 1, 2$, and only if $y^0/Y = z^0/Z$.

The necessity of $y^0/Y = z^0/Z$ follows immediately from Theorem 1, and the sufficiency of $y^0/Y = z_j^0/Z_j$, $j = 1, 2$, is shown in the Appendix. Corollary 1 establishes that portfolio diversification cannot always duplicate return distributions that are attainable through merger. But mergers are not costless, because other return distributions can be acquired only through portfolio diversification. This conclusion follows

immediately from Theorem 1 and is stated as[5]

COROLLARY 2: *Given* **H.1** *and* **H.2**, *portfolios of securities in the separate corporations that do not have the property* $y_1/Y_1 = z_1/Z_1 = y_2/Y_2 = z_2/Z_2$ *yield returns that for some* θ_j *cannot be duplicated through holdings of the conglomerate's securities.*

Corollaries 1 and 2, taken together, identify potential benefits and costs to investors from conglomeration; however the benefits can occur only if there is default risk on the separate corporations' bonds. This conclusion is stated as Theorem 2. It is proved using two hypotheses: **H.3** $pr\{\theta_j < \theta_j^*\} = 0, j = 1, 2$; and **H.4** there exist at least three ordered pairs of values of the random variables that are not linearly related and are assigned positive probabilities.

THEOREM 2 (The Homemade Mutual Fund Theorem): *Given* **H.3** *and* **H.4**, *each* (z^0, y^0) *yields a return distribution that would be duplicated for all* θ_j *only by* $(\hat{z}_1, \hat{z}_2, \hat{y}_1, \hat{y}_2)$ *satisfying*

$$\sum_j (\hat{y}_j + \hat{z}_j) \leqq \sum_j (y_j^0 + z_j^0)$$

such that $\sum_j \hat{z}_j = z^0$ *and* $\hat{y}_j/Y_j = y^0/Y$, $j = 1, 2$.

The theorem is proved in the Appendix, and its interpretation is straightforward. In the absence of default risk, merger would seem to be undesirable. For each (z^0, y^0), the return distribution from a portfolio of the conglomerate's securities could have been acquired through a diversified portfolio. Any diversified portfolio such that $y_1/Y_1 \neq y_2/Y_2$ does not duplicate a return distribution from a portfolio of the conglomerate's securities. Merger could reduce investors' flexibility of choice.

The arguments in this section have not required that security prices be unaffected by merger or that they change in any specified way. The effect of merger on opportunity sets follows from the conditions of the merger, and the effects on security prices and firms' market values would follow from the way opportunity sets are changed. Although the results establish conditions such that investors would not be indifferent to merger, so that security prices could be affected, the conditions are not informative about preferences for or against it.

II. Cheap Credit as an Explanation of Mergers

Arguments have been made that conglomeration reduces the risk of investments and could thereby lower the cost of funds. I interpret these arguments to mean that conglomeration in the presence of default risk changes return distributions on securities to make them more attractive to risk-averse investors. To relate risk to unambiguous effects on the cost of capital, the appropriate measure of risk should imply that less risky distributions are ranked at least as high as more risky distributions by all concave utility functions. This meaning of relatively less risk is used through the rest of the paper.

The definition of relative risk is distinct from the variance definition, which has been common to arguments that merger reduces risk.[6] Yet the definition is consistent with an intuitive meaning of relative risk, because it implies that one distribution

[5] If the hypothesis in Corollary 2 that $pr\{\theta_j < -1\} = 0$, $j = 1, 2$ is altered to admit negative cash flows for the corporations, then all portfolios of stocks and bonds of the separate corporations yield return distributions that are distinct from any available from portfolios of the conglomerate's securities. This result follows from limited corporate liability. If $\theta_1 < -1$ and $\theta_2 > -1$ such that $-1 < \theta < \theta^*$, an investor who owns a diversified portfolio would receive nothing from the first corporation and would get some return on the bonds of the second. If the corporations merge, the earnings of one of the conglomerate's activities would have to cover accounts payable of the other. Thus even if $z_1/Z_1 = z_2/Z_2$, merger could alter the distribution of returns on a portfolio, if values of $\theta_1 < -1$ are admissible.

[6] Michael Rothchild and Stiglitz show that the partial ordering of distributions based on the definition of relative risk used in this paper can be distinct from orderings based on the variance definition of risk.

is less risky than another only if every risk averter prefers the former. More importantly, unless at least one return distribution attainable only through merger is less risky than all distributions available through portfolio diversification, then the assertion that merger causes cheap credit by reducing risk is a nonsequitur. Unanimous investor support for the merger would not follow from a general model of individual maximizing behavior, and a merger's effects on firms' market values and on the cost of capital would depend on preferences of investors—some who support and others who may oppose the merger.

In the presence of default risk, merger can have an adverse affect on stock yields because returns that would have gone to stockholders of one distinct corporation can go to cover another's debt obligations. Of course, investors may be able to recapture returns by purchasing debt. If they could, then merger need not imply that investors who hold stock in their portfolios must forego returns. Theorem 3 establishes that merger in the presence of default risk must always affect stock yields in a way that some investors could find undesirable, and Theorem 4 states that the undesirable effect could not be offset by trading the conglomerate's stocks for its bonds at the equilibrium rate of exchange.

THEOREM 3: *Given the hypothesis that default risk exists, at least one concave utility function must always exist that would rank the return distribution from a unit of stock of at least one of the two corporations higher than the return distribution from a unit of the conglomerate's stock.*

The proof of the theorem is in the Appendix. The theorem holds because merger must reduce the expected return from the stock of at least one of the two distinct corporations, the corporation with stock that yields the higher expected return. If the stocks of the two corporations yield the same expected return, merger would lower it.

Any investor who held a portfolio $y_i^0 > 0$, $y_k^0 = z_i^0 = z_k^0 = 0$ would get a portfolio $y^0 = y_i^0$, $z^0 = 0$ as a direct consequence of merger. If the expected return from the stock of corporation i were at least as great as the expected return from the stock of corporation k, then merger would reduce the portfolio's expected return. Of course, an investor could trade his endowment of the conglomerate's stock $y^0 = y_i^0$ for bonds at the equilibrium rate of exchange. But even if such exchanges were made, an investor could not acquire a portfolio with a greater expected return than the expected return from the portfolio $y^0 = y_i^0$, $z^0 = 0$. This conclusion is shown in the Appendix and serves as part of the proof of

THEOREM 4: *Given the hypotheses that default risk exists and investors' preferences can be represented by concave utility functions defined over terminal wealth, then at least one utility function must rank the return distribution from a portfolio containing only the stock of one of the distinct corporations higher than any return distribution obtainable from an endowment of $y^0 = y_i^0$ or from subsequent trades at the equilibrium rate of exchange between the conglomerate's stocks and bonds.*

Theorems 3 and 4 establish that if merger does not alter the probability of any θ_j, merger does not necessarily make return distributions less risky. Although the two theorems are sufficient to show that merger does not necessarily provide mutual benefits by reducing risk for investors in either of the distinct corporations, the theorems lead to a conclusion that is in a sense unfortunate. They imply that empirical models that have been used to relate merger activity to readily quantifiable data cannot be deduced from general models of individual maximizing behavior. Constraints must be placed on either the set of admissable preferences or on the set of joint density functions; but Theorems 3 and 4 suggest that the search for reasonable a priori constraints on joint density functions will probably not be fruitful, since the theorems are general to any joint density function.

Theorems 3 and 4 suggest that the nature of investors' preferences may have to be in-

voked to explain merger. But inferences about the causes of mergers may be possible from data on merger activity; however, the inferences may require detailed data about the portfolios of investors who have significant control over firms with aggressive acquisition policies. Such data would reveal information about those investors' preferences and possibly could be used to make inferences about mergers' benefits.

Since merger must reduce the expected return on the stock of at least one of the merging corporations, a logical source of benefits is conglomeration's effect on bond returns. This observation emphasizes the importance of disaggregated data on portfolio composition and on the cost of debt finance, since stockholders who hold significant amounts of bonds in their portfolios could benefit from higher expected portfolio returns or capital gains and all stockholders could benefit from lower costs of debt finance.

An example of merger's effect on bond returns can be readily identified. In order to abstract from the effects of different conditional density functions on rates of return and different corporate financial structures, assume that the distinct firms have identical conditional density functions on rates of return and that they have financed capital expenditures by issuing debt and equity in the same proportions. An additional assumption serves to identify benefits for bondholders. Exclude negative cash flows. Since bondholders have first claim on returns only after the costs of production have been covered, then even though the expected return on stocks would decrease in the event of conglomeration, the returns which some stockholders might forego could go to cover accounts payable rather than to investors in corporate debt.

Since the two firms are essentially identical, their stocks will yield the same expected return, and merger will reduce it.[7] Since negative cash flows are excluded, any reduction in stocks' expected returns will be reflected in increased expected returns on bonds. Furthermore, merger increases the expected return on any portfolio of the distinct corporations' securities that contains a larger proportion of bonds to equity than the proportion of bonds to equity issued by the firms.[8]

The example suggests some truth to the argument that if firms merge, bondholders may benefit and the conglomerate may be able to offer more attractive bonds than those of the distinct corporations. Since there is some truth to the argument, the effects of conglomeration on bond prices may make a sensible empirical study, although the support for the cheap credit argument is tenuous. Examples can be constructed such that merger could lower the expected return and increase the variance on each dollar of redeemable bond value offered by one of the distinct corporations. Besides, as Theorem 4 indicated, stockholders may dislike benefits conferred on bondholders.

III. Conclusion

Conglomerate mergers do occur, and special cases can be found such that mergers would be preferred by at least some investors; but such cases would not hold for

---

[7]The proof of Theorem 3 in the Appendix is done under the most general conditions, θ_i, $\theta_k \in (-\infty, \infty)$. The theorem holds for the special case of $pr\{\theta_j < -1\} = 0$. Merger would still reduce the expected return on the stock of at least one of the corporations; however a proof using -1 as the lower limit of integration requires an additional limit of integration on θ_k to assure $g(\theta_k) \geq -1$. If negative cash flows are excluded, $E(e)$ in (A7) becomes

$$\int_{-1}^{b} \int_{g(\theta_k)}^{\infty} ef(\theta_i, \theta_k)\, d\theta_k d\theta_i + \int_{b}^{\infty} \int_{-1}^{\infty} ef(\theta_i, \theta_k)\, d\theta_k d\theta_i$$

where $b = (\alpha_i + \theta^*)\alpha_k^{-1}$. A trivial change from $-\infty$ to -1 is made in the lower limit of integration in (A6). The proof of the special case of Theorem 3 is direct.

[8]Proofs of the propositions about changes in expected bond and expected portfolio returns as a consequence of the merger of essentially identical firms have been omitted, because the equations needed in the proofs are lengthy and the propositions seem intuitive.

all concave utility functions and all possible portfolios. Yet inferences about the causes of merger activity may still be possible, because merger can have desirable effects on bonds. There may be some validity to the cheap credit argument, if it is not overstated. It would not necessarily hold for a combination of any two firms with unrelated business activities.

Firms often spend substantial search costs in the process of finding merger partners. Considerable nonmarker information is exchanged in the negotiation of a merger, and not all proposed mergers are consummated. This suggests that corporate owners and managers must search for merger partners with debt to equity ratios and return distributions to their liking. Even if they find suitable partners, those partners may be unwilling, because the willingness to merge would not be independent of the compositions of optimal portfolios of that minority of most corporations' owners who control them.

The model includes only one period of returns. In a two-period model, merger could affect the probability of returns in ways not accounted for in a one-period model, because returns of one firm could support the debts of the other to avoid bankruptcy and maintain production in the next period. Even with the added generality of a two-period model, one could still not show that merger would in general be preferred by all investors with concave preferences to maintaining investments in at least one of the distinct corporations. Examples could be constructed such that merger would reduce the expected present discounted returns on the stock of at least one of the distinct corporations.

APPENDIX

PROOF of Theorem 1:

Statements (5) and (12) imply that if $W = W'$ for all θ_j, then

$$\text{(A1)} \qquad \sum_j R_j = R \quad \text{for } \theta_j < \theta_j^* \qquad j = 1,2$$

$$\text{(A2)} \qquad R_i + R_k^* = R \quad \text{for } \theta_i < \theta_i^* \text{ and } \theta_k \geqq \theta_k^* \text{ which imply } \theta < \theta^* \qquad i,k = 1,2 \quad i \neq k$$

$$\text{(A3)} \qquad R_i + R_k^* = R^* \text{ for } \theta_i < \theta_i^* \text{ and } \theta_k \geqq \theta_k^* \text{ which imply } \theta \geqq \theta^* \qquad i,k = 1,2 \quad i \neq k$$

$$\text{(A4)} \qquad \sum_j R_j^* = R^* \text{ for } \theta_j \geqq \theta_j^* \qquad j = 1,2$$

Furthermore, since $\sum_j Z_j(1 + r_j) = Z(1 + r)$, if $z_j/Z_j = z/Z$, then

$$\text{(A5)} \qquad \sum_j z_j(1 + r_j) = z(1 + r)$$

Inspection of (3), (4), (10), and (11) reveals that R_j, R_j^*, R, and R^* are linear functions of θ_j. If for each of the ranges of random variables listed in (12) and again in (A1)-(A4), there exists three ordered pairs of values of the random variables that are each assigned positive probability, then (A1)-(A4) could hold as equalities for all θ_j only if the coefficients on θ_j are the same, or in other words only if $y/Y = z/Z = y_j/Y_j = z_j/Z_j$. The proof of sufficiency is straightforward. Inspection of (3) and (10) reveals that $z/Z = z_j/Z_j$, $j = 1, 2$, is sufficient for (A1) to hold, and (4) and (11) reveal that if (A5) holds, $y/Y = y_j/Y_j$, $j = 1, 2$, is sufficient for (A4). Furthermore, if $y_k/Y_k = z_k/Z_k$, then (4) reduces to $R_k^* = S_k((1 + \theta_k) y_k/Y_k)$, and then (3), (4), and (10) imply that (A2) holds if $z/Z = y_k/Y_k = z_i/Z_i$. Finally, $y/Y = z_i/Z_i = z_k/Z_k$ and (A5) yield the result that (11) reduces to $R^* = (S_k(1 + \theta_k) + S_i(1 + \theta_i))\, y/Y$, and since (4) reduces to $R_k^* = S_k(1 + \theta_k)y_k/Y_k$, then (3), (4), and (11) imply that (A3) holds if $y/Y = y_k/Y_k = z_i/Z_i$.

PROOF of Corollary 1:

Theorem 1 states that $y^0/Y = z^0/Z$ must hold if there is to be a corresponding portfolio $\hat{y}_j$ and $\hat{z}_j$ of the separate corporations' securities that yields the same returns for all θ_j as the returns from y^0 and z^0. If $y^0/Y =$

z_j^0/Z_j, $j = 1, 2$, then $y^0/Y = z^0/Z$, since $z^0(1 + r) = \sum_j z_j^0(1 + r_j)$ and $Z(1 + r) = \sum_j Z_j(1 + r_j)$; therefore, the necessary condition is satisfied. Furthermore, $\hat{y}_j$ and $\hat{z}_j$ will exist that satisfy the expenditure constraint. Since $Y = \sum_j Y_j$, then $\hat{y}_j$ must exist that simultaneously satisfy $\hat{y}_j/Y_j = y^0/Y$, $j = 1, 2$, and $\sum_j \hat{y}_j = y^0 = \sum_j y_j^0$. Finally, since $Z(1 + r) = \sum_j Z_j(1 + r_j)$ and $z^0 \cdot (1 + r) = \sum_j z_j^0(1 + r_j)$, if $z_1^0/Z_1 = Z_2^0/Z_2$, then $\hat{z}_j = z_j^0$ assures $\hat{z}_j/Z_j = z_j^0/Z_j = z^0/Z$. Consequently, $\sum_j(\hat{y}_j + \hat{z}_j) = \sum_j(y_j^0 + z_j^0)$ and $y^0/Y = z^0/Z = \hat{y}_j/Y_j = \hat{z}_j/Z_j$ for $j = 1, 2$, and by Theorem 1, the portfolios y^0, z^0, and $\hat{y}_j$, $\hat{z}_j$ must have the same returns for all θ_j.

The condition $y^0/Y = z^0/Z$ is necessary but not sufficient. If $z_1^0/Z_1 \neq z_2^0/Z_2$, then even if $y^0/Y = z^0/Z$, the corresponding portfolio $\hat{y}_j$, $\hat{z}_j$ that satisfies the expenditure constraint may not exist. If $r_1 \neq r_2$, then the $\hat{z}_j$ such that $\hat{z}_j/Z_j = z^0/Z$, $j = 1, 2$, could be such that $\sum_j \hat{z}_j > \sum_j z_j^0$, which would imply $\sum_j(\hat{y}_j + \hat{z}_j) > \sum_j(y_j^0 + z_j^0)$.

PROOF of Theorem 2:

Without default risk on bonds, returns from a diversified portfolio and returns from a portfolio of the conglomerate's securities would be the same if and only if (A4) holds. Since there exists at least three ordered pairs of values of the random variables that are assigned positive probabilities and are not linearly related, (A4) could hold only if $\hat{y}_j/Y_j = y^0/Y$. Since $Y = \sum_j Y_j$, the $\hat{y}_j$ are unique and satisfy $\sum_j \hat{y}_j = y^0 = \sum_j y_j^0$. Furthermore, (4) and (11) imply that if $\hat{y}_j/Y_j = y^0/Y$, then (A4) will hold only if $\sum_j \hat{z}_j(1 + r_j) = z^0(1 + r)$. Statement (9) implies that, since there is no default risk on the distinct corporations bonds, there is no default risk on the conglomerate's bonds. Consequently, $r_j = r = \bar{r}$, the risk-free rate of interest. Any $\hat{z}_j$ such that $\sum_j z_j = z^0$ would satisfy $\sum_j \hat{z}_j(1 + r_j) = z^0(1 + r)$ and $\sum_j \hat{z}_j = \sum_j z_j^0$.

PROOF of Theorem 3:

Prior to merger, a unit of stock of either corporation would return $e_j Y_j^{-1}$, where $e_j \equiv (1 + \theta_j)S_j - Z_j(1 + r_j)$ if $\theta_j \geq \theta_j^*$ and $e_j \equiv 0$ if $\theta_j < \theta_j^*$, $j = 1, 2$.

The expected return on a unit of stock would be $E(e_j)Y_j^{-1}$, and for one of the corporations, corporation i, this is

$$\text{(A6)}\quad E(e_i)Y_i^{-1} = Y_i^{-1}\int_{-\infty}^{\infty}\int_{\theta_i^*}^{\infty} e_i f(\theta_k, \theta_i)\, d\theta_k d\theta_i \qquad \text{for } i, k = 1, 2; i \neq k$$

where f is the joint density function.

A unit of the conglomerate's stock yields $e(Y_i + Y_k)^{-1}$ where $e \equiv \sum_j((1 + \theta_j)S_j - Z_j(1 + r_j))$ if $\theta \geq \theta^*$ and $e \equiv 0$ if $\theta < \theta^*$. The expected return is

$$\text{(A7)}\quad E(e)(Y_i + Y_k)^{-1} = (Y_i + Y_k)^{-1} \int_{-\infty}^{\infty}\int_{g(\theta_k)}^{\infty} ef(\theta_k, \theta_i)\, d\theta_k d\theta_i$$

and $g(\theta_k) = \theta_i^* + \alpha_k\alpha_i^{-1}(\theta_k^* - \theta_k)$. This limit of integration is derived from (9), the definition of θ^*, and the observation that either $\sum_j \alpha_j\theta_j \geq \theta^*$ or the stock returns nothing.

Since the distinct corporations are assumed to have sold stock with positive expected returns, $(E(e_i) = E(e_k))(Y_i + Y_k)^{-1} \leq \max\{E(e_i)Y_i^{-1}, E(e_k)Y_k^{-1}\}$. If $E(e) < E(e_i) + E(e_k)$, then

$$\text{(A8)}\quad E(e)(Y_i + Y_k)^{-1} < \max\{E(e_i)Y_i^{-1}, E(e_k)Y_k^{-1}\}$$

Statement (A8) holds because

$$\text{(A9)}\quad \int_{\theta_k^*}^{\infty}\int_{\theta_i^*}^{\infty} (e_i + e_k)f(\theta_k, \theta_i)\, d\theta_k d\theta_i + \int_{-\infty}^{\theta_k^*}\int_{\theta_i^*}^{\infty} e_i f(\theta_k, \theta_i)\, d\theta_k d\theta_i + \int_{\theta_k^*}^{\infty}\int_{-\infty}^{\theta_i^*} e_k f(\theta_k, \theta_i)\, d\theta_k d\theta_i$$
$$> \int_{\theta_k^*}^{\infty}\int_{\theta_i^*}^{\infty} ef(\theta_k, \theta_i)\, d\theta_k d\theta_i + \int_{-\infty}^{\theta_k^*}\int_{g(\theta_k)}^{\infty} ef(\theta_k, \theta_i)\, d\theta_k d\theta_i + \int_{\theta_k^*}^{\infty}\int_{g(\theta_k)}^{\theta_i^*} ef(\theta_k, \theta_i)\, d\theta_k d\theta_i$$

The sum of terms on the left in (A9) is simply $E(e_i) + E(e_k)$, and the sum of terms on the right is $E(e)$. The first term on the left in (A9) equals the first term on the right, because $e_i + e_k = e$ for $\theta_i \geqq \theta_i^*$ and $\theta_k \geqq \theta_k^*$. The second term on the left exceeds the second term on the right, because $g(\theta_k)$ and e are everywhere linear in θ_k and $g(\theta_k) > \theta_i^*$ and $e < e_i$ for $\theta_k \in (-\infty, \theta_k^*)$. Likewise, the third term on the left exceeds the third term on the right. Statement (A9) holds as a strict inequality so long as there is default risk on the bonds of at least one of the two merging corporations. The implication of (A8) is straightforward. If $E(e_i)Y_i^{-1} \geqq E(e_k)Y_k^{-1}$, there must exist at least one concave utility function that would rank a portfolio with the composition $y_i > 0$, $y_k = z_k = z_i = 0$ higher than the return distribution from a merger that generates a portfolio of $y = y_i, z = 0$.

PROOF of Theorem 4:

Merger can cause the price of a unit of stock to change; therefore denote the market price of a unit of the conglomerate's stock as $p \gtreqless 1$. Likewise, the redemption value of bonds with a market value of one dollar is $1 + r \gtreqless 1 + r_j$. If an investor trades a unit of the conglomerate's stock for bonds, he would receive pZ^{-1} of total returns available to bondholders in the event of default and $p(1 + r)$ in the event of repayment. Statements (1) and (6)–(11) are used to find the expected return on bonds that would be received in exchange for a unit of stock. The expected return is

$$\text{(A10)}\quad pZ^{-1}\int_{-\infty}^{\infty}\int_{h(\theta_k)}^{g(\theta_k)}((1+\theta_i)S_i + (1+\theta_k)S_k)f(\theta_k,\theta_i)\,d\theta_k d\theta_i + p(1+r)\int_{-\infty}^{\infty}\int_{g(\theta_k)}^{\infty}f(\theta_k,\theta_i)\,d\theta_k d\theta_i$$

where $h(\theta_k) = -(1 + \alpha_k\theta_k)\alpha_i^{-1}$. This lower limit of integration in the first term in (A10) is derived from the observation that either $\sum_j \alpha_j\theta_j \geqq -1$ or the conglomerate's bonds return nothing, because it has a negative cash flow. The first term in (A10) is the expected return on bonds in the event of default and the second term is the expected return in the event of repayment.

Since p is an equilibrium price of stocks, then at least one investor with a utility function U defined over terminal wealth must choose to hold some stocks in his portfolio. The proof of the theorem will be accomplished by showing that no risk-averse or risk-neutral investor would choose to hold the conglomerate's stock if p were so large that the expected return on a unit of stock is less than the expected return on the bonds that could be received in exchange for it.

If W_0 is the investor's initial wealth endowment and x is the amount invested in riskless bonds, then the wealth constraint is

$$\text{(A11)}\quad W_0 = x + z + py$$

The investor's expected utility is

$$\text{(A12)}\quad V = U(N)pr\{\theta \leq -1\} + \int_{-\infty}^{\infty}\int_{h(\theta_k)}^{g(\theta_k)}U(R+N)f(\theta_k,\theta_i)\,d\theta_k d\theta_i + \int_{-\infty}^{\infty}\int_{g(\theta_k)}^{\infty}U(R^*+N)f(\theta_k,\theta_i)\,d\theta_k d\theta_i$$

where N is nonrandom income and includes the return on riskless bonds. The Lagrange function is

$$\text{(A13)}\quad L = V + \lambda(W_0 - x - z - py)$$

and the Kuhn-Tucker necessary conditions imply that an investor would choose to hold a portfolio satisfying $y > 0$ and $z \geqq 0$ only if

$$\text{(A14)}\quad (Y_i + Y_k)^{-1}\cdot\int_{-\infty}^{\infty}\int_{g(\theta_k)}^{\infty}eU'(R^*+N)f(\theta_k,\theta_i)\,d\theta_k d\theta_i \geqq pZ^{-1}\int_{-\infty}^{\infty}\int_{h(\theta_k)}^{g(\theta_k)}U'(R+N)\cdot((1+\theta_i)S_i + (1+\theta_k)S_k)f(\theta_k,\theta_i)\,d\theta_k d\theta_i + p(1+r)\int_{-\infty}^{\infty}\int_{g(\theta_k)}^{\infty}U'(R^*+N)\cdot f(\theta_k,\theta_i)\,d\theta_k d\theta_i$$

Obviously, if U is linear, (A14) could hold only if (A7) is at least as great as (A10). If U is strictly concave, then (A14) could hold only if (A7) exceeds (A10), because e is increasing in $\theta_k \in (-\infty, \infty)$ and $\theta_l \in [g(\theta_k), \infty)$, and R^* for all $\theta_k \in (-\infty, \infty)$ and $\theta_l \in [g(\theta_k, \infty)$ is at least as great as R for all $\theta_k \in (-\infty, \infty)$ and $\theta_l \in [h(\theta_k), g(\theta_k)]$, since R^* represents returns in the event of complete repayment of corporate debt, and R represents returns in the event of default. Consequently, if U is concave, (A14) could hold only if the expected return on a unit of stock, represented by (A7), were at least as great as the expected return on the bonds, represented in (A10), that would be received in exchange for a unit of stock.

If p is to be an equilibrium price of stock, then it could not be so high that a unit of stock could be exchanged for bonds with a larger expected return. If p were that high, it could not be an equilibrium price, since no risk-averse or risk-neutral investor would choose to hold stocks in an optimal portfolio. Consequently any investor who receives $y^0 = y_l^0$ of the conglomerate's shares as part of his wealth endowment could not increase the expected return on his portfolio by trading stocks for bonds at the equilibrium p. This result and (A8) are sufficient for the theorem.

REFERENCES

M. Adelman, "Antitrust Problems: The Anti-merger Act, 1950-1960," *Amer. Econ. Rev. Proc.*, May 1961, *51*, 236-44.

D. Baron, "Default Risk and the Modigliani-Miller Theorem: A Synthesis," *Amer. Econ. Rev.*, Mar. 1976, *66*, 204-12.

D. Cass and J. Stiglitz, "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocations," *J. Econ. Theory*, June 1970, *2*, 122-60.

M. Gort, "An Economic Disturbance Theory of Mergers," *Quart. J. Econ.*, Nov. 1969, *83*, 624-42.

K. Hagen, "Default Risk, Homemade Leverage, and the Modigliani-Miller Theorem: Note," *Amer. Econ. Rev.*, Mar. 1976, *66*, 199-203.

F. Modigliani and M. Miller, "The Cost of Capital, Corporate Finance, and the Theory of Investment," *Amer. Econ. Rev.*, June 1958, *48*, 261-97.

M. Rothchild and J. Stiglitz, "Increasing Risk I: A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225-43.

V. Smith, "Corporate Financial Theory," *Quart. J. Econ.*, Aug. 1970, *84*, 451-71.

J. Stiglitz, "A Re-examination of the Modigliani-Miller Theorem," *Amer. Econ. Rev.*, Dec. 1969, *59*, 784-93.

Success Indicators in the Soviet Union: The Problem of Incentives and Efficient Allocations

By MARTIN LOEB AND WESLEY A. MAGAT*

> ... [T]he correct combination of the interests of the controlling and the controlled levels of production is one of the most important tasks of an optimally controlled economy. It is quite possible (and even highly probable!) that the liquidation of the striving of the lower levels to hide their productive possibilities, the orientation of the interests of the masses to the search for new, better variants of production and many other consequences of such a combination, at the present time conceals bigger reserves for the growth of the socialist economy, than the use of mathematical programming with the preservation of the former relations between the controlling and the controlled levels of the economy.
>
> *Viktor V. Novozhilov* [p. 32][1]

As Novozhilov recognized in 1969, elimination of the incentives for Soviet managers to act contrary to the interests of society would yield significant gains in the value of output produced by Soviet enterprises. Realizing the importance of incentives for motivating behavior which is consistent with plans, Soviet planners have long struggled with the problem of designing good success indicators.[2] Success indicators should induce enterprises to act efficiently in carrying out state plans. They should also motivate enterprises to send accurate forecasts to the Central Planning Bureau (*CPB*) so that socially desirable plans may be constructed and capital efficiently allocated.

Forecasts are particularly important in a centralized economy because the *CPB* cannot be expected to possess detailed knowledge of the enterprises' production functions. Nove presents a convincing argument which explains why enterprises are constantly presented with a range of alternative decisions, that is, why the *CPB* is incapable of centrally planning all aspects of the production process. He also provides some examples of how poor success indicators provide incentives for enterprises to produce inefficiently, or to otherwise act contrary to the best interests of the *CPB*, and thus the Soviet state.

As Joseph Berliner has explained (pp. 401, 402), Soviet planners cannot reasonably expect managers to act as *Homo Sovieticus*, the party man who carries out all directives and makes all decisions in the best interests of his country. *Homo Economicus*, the manager who acts in his own best interests, provides a more reliable model for studying the behavior of the Soviet manager, for his operating rules are derived primarily from the incentive structure and may differ considerably from the formal decision rules. As an example, the ratchet problem which has plagued Soviet planning can only be explained by reference to an incentive structure which penalizes managers by assigning them higher plans in subsequent years if they overfulfill their current year's plans by too great a margin. As David Granick concludes,

> Soviet leaders have viewed intermediate and lower-level managers as "economic men"—much as top decision makers in capitalist firms are viewed in orthodox neoclassical economic theory. They have perceived their own problem as being that of creating a combined incentive and de-

*Assistant professor, department of economics and business, North Carolina State University, and assistant professor, Graduate School of Business Administration, Duke University, respectively.

[1]This quote was taken from Michael Ellman (1973b, p. 40).

[2]See Alec Nove.

cision-rule system which would lead such managers, in their own personal and narrow self-interest, to act in the fashion desired by the central policy makers. [p. 37]

Our work takes this same view in analyzing the "bonus-maximizing" behavior of Soviet managers. We do not distinguish between the behavior of the enterprise and that of its managers, since bonuses are the prime source of income over which managers have control and bonuses are tied closely to the enterprise bonus fund. In addition, the earnings of the managerial and professional staff are also dependent upon the size of the bonus fund, so bonus maximization helps to foster a pleasant working atmosphere and to retain reliable staff members.[3]

Western economists have recently begun to analyze the incentive properties of success indicators employed in the Soviet Union. The growing literature on resource allocation mechanisms[4] has also spurred renewed interest in the design and analysis of new success indicators with desirable incentive properties. Ellman (1971, 1973a, b) provides an algebraic characterization of the bonus formulas incorporated in the 1965 Soviet economic reform. Martin Weitzman offers a formulation of the success indicators utilized in the 1971 reform and a careful analysis of their incentive properties, both in the certainty and uncertainty cases. On the design side, Liang-Shing Fan has proposed a success indicator to motivate desirable behavior on the part of Soviet enterprise managers which also possesses desirable incentive properties.

This paper accomplishes three main purposes. First, we show that the reform of 1971, in so far as it relates to the static incentive properties of enterprise success indicators, is merely cosmetic and has the same properties as those given by the 1965 reform. The Weitzman indicators are shown to be only a slight generalization of the Ellman indicators. Furthermore, we observe that Fan's proposal is not new, since the Fan indicators are a subset of the Ellman indicators (and, hence, also a subset of the Weitzman indicators).

Our second purpose is to show that the Fan, Ellman, and Weitzman success indicators actually possess *undesirable* incentive properties. While Fan, Ellman, and Weitzman all recognize the desirability of motivating accurate forecasts, their models do not explain how the *CPB* uses these forecasts. Their analyses implicitly assume that enterprises ignore the effects of forecasts on *CPB* allocations to the enterprises. Our model recognizes that the *CPB* uses forecasts to make allocations. We also allow enterprises to take this knowledge into account when sending forecasts to the *CPB*. Under such circumstances, we show that using the success indicators studied by Fan, Ellman, and Weitzman, enterprises can individually gain by transmitting inaccurate forecasts, to the detriment of society as a whole.

The third purpose of this paper is to present a new success indicator which motivates accurate forecasts and efficient behavior in the context of our more general model.

In Section I we show that the set of Fan success indicators is a subset of the Ellman indicators, which in turn comprise a subset of the Weitzman indicators. Section II contains a description of the problem which arises when the *CPB* uses forecasts to allocate capital to the enterprises. We define the incentive problem in terms of a game in which enterprises are motivated by success indicators to play strategies consisting of forecasts and operating decisions. Section III provides an example to demonstrate that the Fan, Ellman, and Weitzman indicators actually encourage the transmittal of biased forecasts. Section IV contains a presentation of a new success indicator designed to overcome the problems associated with the Weitzman, Ellman, and Fan indicators.

I. The Weitzman, Ellman, and Fan Success Indicators

Weitzman has supplied a major contribution to the study of Soviet incentive

[3] See Granick, pp. 37–41.
[4] For a review of this literature, see Leonid Hurwicz.

systems by formally modeling the success indicators imbedded in the recent 1971 Soviet economic reform.[5] He is careful to explain that his "aim is to focus directly on the *analytical essence* of the new reward structure" (p. 252, emphasis added). Our work follows in the spirit of his paper, since we believe that any complex reward system cannot be expected to motivate socially desirable behavior unless, when reduced to its barest essentials, it possesses the necessary incentive properties.

We recognize that there are several enterprise incentive funds (each based on more than one variable, such as sales or profit rate), that managers receive additional bonuses related to such activities as development and assimilation of new technology, and that enterprises face a considerable amount of administrative uncertainty. Further, a success indicator which possesses desirable incentive properties within the context of a certainty model that incorporates only a single bonus fund may not actually induce the socially desired behavior in this more complex environment. However, we cannot expect to understand the effects of complicated reward systems until we first understand the properties of the simpler incentive systems on which they are based.

Weitzman describes a three-stage process. In the preliminary phase, planners assign each enterprise a tentative target π° and a tentative bonus $\bar{B}$, where the target may reflect profit rate, value of output, or labor productivity. The enterprises are also assigned bonus and penalty coefficients, α, β, and γ, where $0 < \alpha < \beta < \gamma$. In the planning phase, each enterprise must choose a plan target π^F, which could differ from the tentative target π°. If the enterprise actually produced the targeted amount π^F, then the enterprise would receive the planned bonus $\hat{B}$, where

$$(1) \qquad \hat{B} = \bar{B} + \beta(\pi^F - \pi^\circ)$$

In the implementation phase the enterprises produce some amount π^A and are rewarded on the basis of the success indicator S, where

$$(2) \qquad S(\pi^F, \pi^A) = \begin{cases} \hat{B} + \alpha(\pi^A - \pi^F), \text{ if } \pi^A \geq \pi^F \\ \hat{B} - \gamma(\pi^F - \pi^A), \text{ if } \pi^A < \pi^F \end{cases}$$

Weitzman does distinguish between the *static* problem described above and the *dynamic* problem which arises when planners use current performance to revise the coefficients of future success indicators (such as the coefficients α, β, and γ). In the static problem enterprises are assumed to maximize the success indicator S, whereas in the dynamic problem they aim to maximize a time-discounted sum of current and future values of S. The 1971 reform was most significant in that it reduced the size of the ratchet effect which is characteristic of the dynamic problem, for it froze the fixed targets (π° and $\bar{B}$) and the coefficients (α, β, and γ) for the entire ninth five-year plan (1971 75). Thus managers were no longer faced with a penalty for excessive overfulfillment of plans, at least in the first three or four years of the five-year planning period. We follow Weitzman in considering only the static problem.

Using equation (1), we can rewrite (2) as

$$(3) \qquad S(\pi^F, \pi^A) = \begin{cases} \bar{B} + \beta(\pi^F - \pi^\circ) + \alpha(\pi^A - \pi^F), \text{ if } \pi^A \geq \pi^F \\ \bar{B} + \beta(\pi^F - \pi^\circ) - \gamma(\pi^F - \pi^A), \text{ if } \pi^A < \pi^F \end{cases}$$

Now define $\mathbf{S}^W$ as the set of all Weitzman success indicators. That is $\mathbf{S}^W$ is the set of all $S(\pi^F, \pi^A)$, as in (3), with $\bar{B} \geq 0$, $\pi^\circ \geq 0$, and $0 < \alpha < \beta < \gamma$.

When the forecast π^F affects S, but not the actual amount selected π^A, the Weitzman indicators possess desirable incentive properties. As shown by Weitzman, in the case of perfect certainty about production and market possibilities, enterprises are motivated to send truthful forecasts. Given these forecasts, they also have incentives to increase actual performance π^A as much as possible.

[5] See Weitzman for a more detailed description of the success indicator.

Next consider the subset S^L of S^W defined by setting $\bar{B} = \pi^\circ = 0$ (i.e., eliminating the preliminary phase of the three-phase planning process). A member of S^L is, therefore, written as

$$(4)\quad S(\pi^F, \pi^A) = \begin{cases} \beta\pi^F + \beta(1 - \epsilon_1)(\pi^A - \pi^F), \text{if } \pi^A \geq \pi^F \\ \beta\pi^F - \beta(1 + \epsilon_2)(\pi^F - \pi^A), \text{if } \pi^A < \pi^F \end{cases}$$

where ϵ_1 and ϵ_2 are defined by $\epsilon_1 = 1 - \alpha/\beta$ and $\epsilon_2 = \gamma/\beta - 1$, with $0 < \epsilon_1 < 1$ and $\epsilon_2 > 0$. Equation (4) can be rewritten as

$$(5)\quad S(\pi^F, \pi^A) = \beta\pi^F - k\beta(\pi^F - \pi^A)$$

where $k = (1 - \epsilon_1)$ if $\pi^A \geq \pi^F$ and $k = (1 + \epsilon_2)$ if $\pi^A < \pi^F$. Note that equation (5) is the Ellman indicator (represented by equation (3) in his 1973 paper).[6] Ellman presents (5) as a formalization of the success indicators employed by the Soviet planners in the 1965 reform. We have shown that S^L is a subset S^W, so that the 1971 reform indicators represent only a slight generalization of the 1965 reform indicators.

Even this, however, overstates the contribution of the 1971 reform. Success indicators are merely used as a *basis* on which to reward enterprises. Hence, a monotonic transformation of an indicator belongs to the same equivalence class, that is, has the same incentive properties. The Weitzman indicators merely add $\bar{B}$ and $\beta\pi^\circ$ to the Ellman indicators (a monotonic transformation) and replace the *CPB*'s control variables, ϵ_1, ϵ_2, and β, with α, β, and γ. Thus the 1965 and 1971 success indicators are essentially the same; the 1971 reform does not represent a significant step in solving the static incentive problem.[7]

Consider the subset S^F of S^L defined by setting $\epsilon_1 = \epsilon_2 = \epsilon$, where $0 < \epsilon < 1$. Substituting into (4), we can represent a member of S^F as

$$(6)\quad S(\pi^F, \pi^A) = \begin{cases} \beta\pi^F + \beta(1 - \epsilon)(\pi^A - \pi^F), \text{if } \pi^A \geq \pi^F \\ \beta\pi^F - \beta(1 + \epsilon)(\pi^F - \pi^A), \text{if } \pi^A < \pi^F \end{cases}$$

Equation (6) can be written as

$$(7)\quad S(\pi^F, \pi^A) = \beta[\pi^F + (\pi^A - \pi^F) - \epsilon\,|\pi^A - \pi^F|]$$

or

$$(8)\quad S(\pi^F, \pi^A) = \beta[\pi^F + (1 + \epsilon \cdot sgn(\pi^F - \pi^A)) \cdot (\pi^A - \pi^F)]$$

When π^A and π^F are interpreted to mean actual and forecasted *profit*, then equation (8) represents the success indicator given by Fan.[8] Thus, his proposed indicator is merely a special case of ones already in use.

For future reference the Fan indicator is further simplified by rewriting (7) as:[9]

$$(9)\quad S(\pi^F, \pi^A) = \beta[\pi^A - \epsilon\,|\pi^F - \pi^A|]$$

We note that this indicator has been studied elsewhere in the context of management performance evaluation in a firm.[10]

II. Incentives and Allocations

A major problem with the theory of Soviet success indicators is its failure to consider the uses of enterprise forecasts. Weitzman, for example, argues that planning is needed for coordination and that forecasts are indispensible for tightly coordinated plans; however, he does not explain how these forecasts are to be used in coordinating the activities of several enterprises. In the (unusual) special case when the information (forecasts) received by the *CPB* is used for decisions which do *not*

[6]Note that in claiming (5) to be the Ellman indicator, we are assuming that the *CPB* uses the *enterprise* forecast π^F as *its* target plan for rewarding the enterprise.

[7]As we mentioned earlier, the 1971 reform is significant in that it did partially resolve the *dynamic* incentive problem.

[8]Fan does not require that $\epsilon < 1$; however, his incentive structure only motivates profit maximization when ϵ is less than one. We also note that while Fan was aware of Ellman's work, he did not explicitly relate his success indicator to those employed in the Soviet Union. His paper is quite confusing on this point.

[9]After this paper was essentially completed, a paper by John Bonin appeared that also reduces Fan's indicator to this equation and analyzes another variant of the Weitzman indicator.

[10]See Yuji Ijiri et al. and Loeb (1974, 1975).

affect the enterprises' measured performance (π^A), then the Weitzman success indicators do possess some desirable properties, as claimed. They motivate enterprises to report truthful forecasts and to operate as efficiently as possible once these forecasts have been made.

We show in Section III that in the more general case where forecasts are used to make decisions affecting measured performance (for example, the allocation of capital to the enterprises), the Weitzman, Ellman, and Fan indicators motivate enterprises to send biased forecasts.

Before illustrating this weakness of the Weitzman, Ellman, and Fan indicators, we first must explain the incentive problem which arises when the *CPB* makes allocations to the enterprises based on their forecasts.[11] Consider a *CPB* with n enterprises under its control. Let us take the index of performance π^A for each enterprise to be enterprise profit. Assume that enterprises take prices as fixed, that is, each enterprise believes that its forecasts and actual profits will not affect the *CPB*'s determination of prices.[12] We also assume that the *CPB* uses enterprise forecasts to make all allocations of capital to the enterprises.

The profits of the ith enterprise depend upon K_i the amount of capital it is allocated and on L_i a vector of local enterprise decisions.[13] The ith enterprise's profits are denoted by $\bar{\pi}_i^A(K_i, L_i)$, and are gross of any charge for capital. We assume that allocations of capital are made prior to local decisions. To insure that enterprise managers act efficiently in their local operations (i.e., they maximize $\bar{\pi}_i^A(K_i, L_i)$ with respect to L_i) we restrict our attention to success indicators which are increasing in an enterprise's own realized profits. Note that the Weitzman indicators satisfy this requirement.

We may now focus on the functions $\pi_i^A(K_i)$, defined by:

$$(10) \quad \pi_i^A(K_i) = \max_{L_i} \bar{\pi}_i^A(K_i, L_i), \quad i = 1, 2, \ldots, n$$

These functions are assumed to exist for each $K_i \geq 0$ and belong to the set of functions Π, defined as:

$$(11) \quad \Pi = \{\pi : R^+ \rightarrow R \mid \pi(\ ) \text{ is continuous and nondecreasing}\}$$

The *CPB* is assumed to have a fixed quantity $\bar{K}$ of capital to allocate to the n enterprises.[14] It collects contingency forecasts from each enterprise showing the projected profits of that enterprise for various allocations of capital from the *CPB*. The contingency forecasts for the ith enterprise are represented by a function $\pi_i^F(K_i)$. The forecasts $\pi_i^F(\ )$ are also assumed to belong to the set Π.

The *CPB* allocates capital to maximize the sum of the enterprises' forecasted profits. That is the *CPB* selects $\hat{K}_1, \hat{K}_2, \ldots, \hat{K}_n$ such that

$$(12) \quad \hat{K}_1, \hat{K}_2, \ldots, \hat{K}_n \text{ maximize} \sum_{i=1}^{n} \pi_i^F(K_i)$$

subject to $\sum_{i=1}^{n} K_i \leq \bar{K}$ and $K_i \geq 0$, $i = 1, 2, \ldots, n$. Notice that each $\hat{K}_i$ is a function of *all* of the enterprises' forecasts. We therefore use the notation $\hat{K}_i = \hat{K}_i(\pi^F)$, where $\pi^F \equiv (\pi_1^F, \ldots, \pi_n^F)$, and where the arguments of each π_j^F have been suppressed. While the

[11] In what follows, we make the simplifying assumption that the *CPB* imposes no charge on the enterprises for their capital allocation. Joseph Berliner, p. 433, indicates that the enterprises are required to pay to the State Treasury a charge of 15 percent of the value of their fixed and working capital. It can easily be shown that the incentive problem still exists when capital charges are made.

[12] We are also implicitly assuming that the *CPB* selects "correct" prices which account for externality effects that may occur. Under these circumstances, enterprises seeking to maximize profits act in the best interests of society.

[13] While we interpret K_i as the ith enterprise's allocation of capital, it more generally may be interpreted as a vector of resources, "public inputs," or decisions with externalities. See Loeb (1975) and Theodore Groves and Loeb.

[14] Edward Ames has pointed out to us in personal correspondence that the planning board can increase, but in practice not decrease the amount of capital which an enterprise has. Our argument could be modified to account for this constraint by treating $\bar{K}$ as the *increment* in capital available for distribution to the enterprises.

indicated maximum may not be unique, the decision rules $\hat{K}_1(\pi^F)$, $\hat{K}_2(\pi^F)$, $\ldots$, $\hat{K}_n(\pi^F)$ represent a particular maximizer.

The value of the ith enterprise's success indicator depends on its realized profits π_i^A, a real number. However, realized profits are a function of the enterprise's rationed capital, itself a function of the enterprises' contingency forecasts, π^F. For this reason we write enterprise i's success indicator as $S_i(\pi_i^A; \pi^F)$.

As mentioned above, if $S_i(\ )$ is increasing in π_i^A, then the ith enterprise has an incentive to act efficiently in choosing its local decisions. We would also like to choose $S_i(\ )$ so that it motivates accurate forecasts. As a minimal test, we require that if a manager is rewarded on the basis of his success indicators and if he has perfect knowledge of his profit function π_i^A, then he should be motivated to send $\pi_i^A(\ )$ as his forecast.[15] Now the definition of an optimal success indicator can be given.

DEFINITION: The success indicator $\hat{S}_i(\pi_i^A; \pi^F)$ is said to be *optimal* if, and only if:

(a) it is *operationally desirable:* for all π^F belonging to the n-fold Cartesian product of Π and for all real numbers π^1 and π^2, $\pi^1 > \pi^2$ implies $\hat{S}_i(\pi^1; \pi^F) > \hat{S}_i(\pi^2; \pi^F)$;

(b) it is *message desirable:* for all π^F belonging to the n-fold Cartesian product of Π and for all π_i^A belonging to Π, $\hat{S}_i(\pi_i^A[\hat{K}_i(\pi^F/\pi_i^A)]; \pi^F/\pi_i^A) \geq \hat{S}_i(\pi_i^A[\hat{K}_i(\pi^F)]; \pi^F)$, where $\pi^F/\pi_i^A \equiv (\pi_1^F, \ldots, \pi_{i-1}^F, \pi_i^A, \pi_{i+1}^F, \ldots, \pi_n^F)$.

Condition (a) requires that an optimal success indicator be increasing in the enterprise's realized profits. Condition (b) requires that "telling the truth" be a dominant strategy equilibrium for the enterprises. That is, each enterprise can not independently gain by reporting inaccurate forecasts no matter what forecasts the other enterprises send. Thus when it employs a message desirable incentive system the *CPB* has justification for accepting the enterprise forecasts as truthful, avoiding the gaming involved in attempting to correct for forecast bias. Notice that with an optimal indicator each enterprise's evaluation is independent of the local decisions of the other enterprises.

III. Nonoptimality of the Weitzman, Ellman, and Fan Indicators

In this section we show that the success indicators given by Weitzman, Ellman, and Fan do not meet the optimality criteria established in Section II. While their indicators do motivate enterprises to maximize profits (i.e., they are operationally desirable), they encourage enterprises to send biased forecasts (i.e., they are not message desirable). To demonstrate this last weakness, we present a simple counterexample showing that when an enterprise is rewarded on the basis of a Weitzman-Ellman-Fan indicator, the manager has an incentive to send biased forecasts, even if all other managers report truthfully. Thus, sending accurate forecasts is not a dominant strategy equilibrium; it is not even a Nash equilibrium.[16]

Consider a *CPB* which has two units of capital to allocate between two enterprises. Suppose each enterprise's actual profit function is

$$\pi_i^A(K_i) = \sqrt{K_i} \qquad i = 1, 2 \tag{13}$$

Let enterprise 1 be rewarded on the basis of the following success indicator:

$$S_1(\pi_1^A; \pi_1^F, \pi_2^F) = \pi_1^A(\hat{K}_1) - \epsilon \mid \pi_1^F(\hat{K}_1) - \pi_1^A(\hat{K}_1) \mid \tag{14}$$

where $0 < \epsilon < 1$ and $\hat{K}_1 = \hat{K}_1(\pi_1^F, \pi_2^F)$ represents the *CPB*'s allocation to enterprise 1. Recall that, by definition,

[15] Actually, it need only be required to send forecasts with the same marginal efficiency of capital schedule as its actual profit function, since such forecasts will lead to socially optimal allocations.

[16] A Nash equilibrium only requires enterprises to report truthfully when all other enterprises act likewise. Hurwicz uses the term "incentive compatibility" to refer to this property of resource allocation mechanisms which lead to behavioral patterns in which "truthful" reporting constitutes a Nash equilibrium.

(15) $\hat{K}_1, \hat{K}_2$ maximize $\pi_1^F(K_1) + \pi_2^F(K_2)$ subject to $K_1 + K_2 \leq 2, K_1 \geq 0$, and $K_2 \geq 0$

Comparing equation (14) to equation (9), we see that enterprise 1 is being rewarded on the basis of a Fan indicator. As $\mathcal{S}^F \subset \mathcal{S}^E \subset \mathcal{S}^W$, it is also being rewarded on the basis of an Ellman and a Weitzman indicator.

Suppose both enterprises send accurate forecasts, that is, send $\pi_i^F(K_i) = \pi_i^A(K_i)$, $i = 1, 2$. One easily verifies that the *CPB* would then set $\hat{K}_1 = \hat{K}_2 = 1$. The profits of enterprise one would equal 1, as would the value of its success indicator (14).

Now suppose enterprise 2 continues to send accurate forecasts, $\pi_2^F(K_2) = \sqrt{K_2}$, but enterprise 1 sends the biased forecast:

$$(16) \quad \pi_1^F(K_1) = K_1 - (1\tfrac{3}{4} - \sqrt{1\tfrac{3}{4}})$$

Using the decision rules given by (15), the *CPB* would then set $\hat{K}_1 = 1\frac{3}{4}$ and $\hat{K}_2 = \frac{1}{4}$. The profits of enterprise 1 increase from 1 to $\sqrt{1\frac{3}{4}}$, as does the value of the success indicator (14). Thus, enterprise manager 1 individually gains by sending biased forecasts, even if enterprise manager 2 reports truthfully, indicating that the Fan, Ellman, and Weitzman indicators are not optimal.[17]

IV. A Class of Optimal Success Indicators

In this section we present a class of success indicators which meet the criteria for optimality given in Section III. Versions of this scheme were independently discovered by William Vickrey, Groves (1969), and Edward Clarke, and have been put forth as an incentive compatible means of allocating both private and public goods.[18]

Consider the following class of success indicators:

$$(17) \quad \hat{S}_i(\pi_i^A, \pi^F) = \pi_i^A(\hat{K}_i) + \sum_{j \neq i} \pi_j^F(\hat{K}_j) - A_i$$

where $\hat{K}_1, \hat{K}_2, \ldots, \hat{K}_n$ are defined by (12) and $A_i = A_i(\pi_1^F, \ldots, \pi_{i-1}^F, \pi_{i+1}^F, \ldots, \pi_n^F)$ is any real value calculated *independently* of enterprise *i*'s forecast.

Success indicator (17) consists of the *i*th enterprise's realized profits (at its allocated level of capital) plus the *forecasted* profits of all other enterprises at their allocated levels of capital, less an amount calculated independently of enterprise *i*'s forecast. Clearly, this success indicator is operationally desirable. In the Appendix, we prove that it is also message desirable; hence, it is optimal.[19] The message desirability of (17) may be explained by examining the enterprise's problem of selecting a forecast. Enterprise *i* selects a forecast π_i^F to maximize indicator (17), which is equivalent to maximizing

$$(18) \quad \pi_i^A(\hat{K}_i) + \sum_{j \neq i} \pi_j^F(\hat{K}_j)$$

since A_i is independent of π_i^F. The forecast π_i^F affects the indicator only through the capital allocations $\hat{K}_1, \hat{K}_2, \ldots, \hat{K}_n$. When enterprise *i* reports its true profit function $\pi_i^F = \pi_i^A$, then from (12) the *CPB* will choose capital allocations that maximize enterprise *i*'s success indicator. Thus by reporting truthful forecasts, the *i*th enterprise ensures that the *CPB* will act to maximize the *i*th enterprise's own success indicator.

As the term A_i does not depend on enterprise *i*'s forecast, it does not affect individual (noncooperative) behavior. However, for a meaningful interpretation of the success indicator which some planners might believe insures an equitable distribution of rewards across enterprises, we suggest the following definition of A_i:

$$(19) \quad A_i \equiv \max_{(K_1, \ldots, K_{i-1}, K_{i+1}, \ldots, K_n)} \left\{ \sum_{j \neq i} \pi_j^F(K_j) \right\}$$

[17]It is interesting to note that with this particular example, the *CPB* cannot detect that enterprise 1 reported a biased forecast, since

$$\pi^F(1\tfrac{3}{4}) = \pi^A(1\tfrac{3}{4}) = \sqrt{1\tfrac{3}{4}}$$

[18]For a survey of this literature, see Loeb (1976).

[19]This result is well known in the literature surveyed in Loeb (1976). An earlier proof appears in Groves and Loeb.

subject to $\sum_{j \neq i} K_j \leq K$ and $K_j \geq 0, j \neq i$

With A_i defined as in (19), the ith enterprise's success indicator measures the contribution that it makes to total profit. When enterprises respond with accurate forecasts, $\hat{S}_i$ will equal the sum of all enterprises' profits less the maximum profits all others could obtain if enterprise i were to be eliminated.[20] That is, $\hat{S}_i$ represents the opportunity cost of abandoning the ith enterprise.

As the indicators of the form (17) are optimal, sending accurate forecasts is a dominant strategy equilibrium. These success indicators therefore solve all incentive problems of a noncooperative game theoretic nature. In addition, Groves (1976) shows that under suitable regularity conditions sending truthful forecasts forms a unique dominant strategy equilibrium, and Jerry Green and Jean-Jacques Laffont show that every indicator with the dominance property can be written in the form given in (17).

Appendix

We must show that the success indicator given by equation (17) is message desirable, but first we introduce some additional notation. Let $\mathcal{K} \equiv \{(K_1, \ldots, K_n) \mid K_j \geq 0, j = 1, 2, \ldots, n \text{ and } \sum_{j=1}^{n} K_j \leq \bar{K}\}$, $\pi_{-i}^F \equiv (\pi_1^t, \ldots, \pi_{i-1}^F, \pi_{i+1}^t, \ldots, \pi_n^F)$, $\hat{K}_i^1 \equiv \hat{K}_i(\pi^t/\pi_i^A)$ and $\hat{K}_i^2 \equiv \hat{K}_i(\pi^t)$. Then

$$\hat{S}_i(\pi_i^A(\hat{K}_i^1); \pi^F/\pi_i^A) - \hat{S}_i(\pi_i^A(\hat{K}_i^2); \pi^F)$$

$$= \left[\pi_i^A(\hat{K}_i^1) + \sum_{j \neq i} \pi_j^F(\hat{K}_j^1) - A_i(\pi_{-i}^t)\right]$$

$$- \left[\pi_i^A(\hat{K}_i^2) + \sum_{j \neq i} \pi_j^t(\hat{K}_j^2) - A_i(\pi_{-i}^t)\right]$$

$$= \left[\pi_i^A(\hat{K}_i^1) + \sum_{j \neq i} \pi_j^F(\hat{K}_j^1)\right]$$

$$- \left[\pi_i^A(\hat{K}_i^2) + \sum_{j \neq i} \pi_j^F(\hat{K}_j^2)\right]$$

$$= \max_{(K_1, \ldots, K_n) \in \mathcal{K}} \left\{\pi_i^A(K_i) + \sum_{j \neq i} \pi_j^F(K_j)\right\}$$

$$- \left\{\pi_i^A(\hat{K}_i^2) + \sum_{j \neq i} \pi_j^F(\hat{K}_j^2)\right\} \geq 0$$

where the last inequality follows by the definition of a maximum and the fact that $(\hat{K}_1^2, \ldots, \hat{K}_n^2) \in \mathcal{K}$. This completes the proof.

[20]By the phrase "if enterprise i were to be eliminated," we mean that enterprise i receives a zero capital allocation and earns a zero profit.

References

Joseph Berliner, *The Innovation Decision in Soviet Industry*, Cambridge, Mass. 1976.

J. Bonin, "On the Design of Managerial Incentive Structures in a Decentralized Planning Environment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 682–87.

E. Clarke, "Multipart Pricing of Public Goods," *Publ. Choice*, Fall 1971, *11*, 17 33.

Michael Ellman, *Soviet Planning Today: Proposals for an Optimally Functioning Economic System*, London 1971.

——, (1973a) "Bonus Formulae and Soviet Managerial Performance: A Further Comment," *Southern Econ. J.*, Apr. 1973, *39*, 652–53.

——, (1973b) *Planning Problems in U.S.S.R.*, London 1973.

L.-S. Fan, "On the Reward System," *Amer. Econ. Rev.*, Mar. 1975, *65*, 226–29.

D. Granick, "Soviet Introduction of New Technology: A Depiction of the Process," unpublished paper, Stanford Res. Inst., Jan. 1975.

J. Green and J. Laffont, "Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods," *Econometrica*, Mar. 1977, *45*, 427–38.

T. Groves, "The Allocation of Resources Under Uncertainty: The Informational Incentive Roles of Prices and Demand in a Team," Center Res. in Manage. Sci., tech. rep. no. 1, Univ. California-Berkeley, Aug. 1969.

———, "Information, Incentives, and Internalization of Production Externalities," in Steven Lin, ed., *Theory and Measurement of Economic Externalities*, New York 1976.

——— **and M. Loeb,** "Incentives and Public Inputs," *J. Publ. Econ.*, Aug. 1975, *4*, 211–26.

L. Hurwicz, "The Design of Mechanisms for Resource Allocation," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 1–30.

Y. Ijiri, J. Kinard, and F. Putney, "An Integrated Evaluation System for Budget Forecasting and Operating Performance with a Classified Budgeting Bibliography," *J. Accounting Res.*, Apr. 1968, *6*, 1–28.

M. Loeb, "Comments on Budget Forecasting and Operating Performance," *J. Accounting Res.*, Autumn 1974, *12*, 363–66.

———, "Coordination and Informational Incentive Problems in the Multidivisional Firm," unpublished doctoral dissertation, Northwestern Univ. 1975.

———, "Alternative Versions of the Demand-Revealing Process," *Publ. Choice*, Spring 1977, Suppl., *29*, 15–26.

V. Novozhilov, "Khozraschetnaya Sistema Planirovaniya," in *Optimal'noe Planirovanie I Sovershenstvovanie Upravlenie Narodnym Khozyaistvom*, Moscow 1969.

A. Nove, "The Problem of 'Success Indicators' in Soviet Industry," *Economica*, Feb. 1958, *25*, 1–13.

W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J. Finance*, May 1961, *16*, 8–37.

M. Weitzman, "The New Soviet Incentive Model," *Bell J. Econ.*, Spring 1976, *7*, 251–57.

Graduate Students in Economics, 1940–74

By WILLIAM E. SPELLMAN AND D. BRUCE GABRIEL*

Since 1904 an annual list of "Doctoral Dissertations in Political Economy in American Universities and Colleges" has appeared in this *Review*. The characteristics of theses in progress during 1904–05 with respect to subject matter and the author's academic origin were tabulated and published in a series of three articles by Lewis Froman. Our examination employs techniques similar to those used by Froman, although we have chosen to include only Ph.D. recipients during 1940–74 rather than all candidates. The data will trace the production of economists by graduate and undergraduate schools, the shift in area of specialization in dissertations, and the changing status of women in the profession since 1904.

The top twenty institutions in terms of quantitative output of Ph.D.s in economics during 1904–74 are listed in Table 1.[1] Their relative contributions during various subperiods are shown to allow the appraisal of each school's importance over time.[2] It is obvious from Table 1 that an individual school's importance is subject to considerable temporal variability. In the earliest era, 1904–39, Columbia was by far the leading producer of economists. During the next twenty years Harvard, Columbia, Wisconsin, and Chicago jointly dominated the field. In the past fifteen years, however, some new leaders in Ph.D. production have joined the elite, notably, Berkeley, Michigan State, Indiana, and Purdue. In fact, between 1970 and 1974, Berkeley was the most productive source of dissertations in economics. Michigan State, which first became a member of the top forty during the 1960's, exhibits the greatest change in rank with its movement to third place for the period 1970–74. New entries in the top twenty for 1970–74 include Princeton, Oregon, George Washington, and UCLA.

As the composition of the leading Ph.D. schools has changed, so has their degree of monopoly in the production of economists. There has been a substantial decline in the concentration of study at the top twenty and top forty schools. The top forty Ph.D. schools produced 99.2 percent of the candidates during 1904–39, 90.5 percent of the degree recipients during 1940–59, and only 51.9 percent of the Ph.D. recipients during 1970–74.[3] This trend is similarly evident for the relative contribution of the top twenty schools from 1904 to 1974, and may be explained in part by the growing availability of Ph.D. programs over time. The rise of New York University, Michigan, Indiana, and Berkeley to the top ten in place of Cornell, Yale, Illinois, and Johns Hopkins in comparing the 1904–39 and 1940–74 periods is demonstrative of the democratization of Ph.D. study in economics and the mobility of graduate programs in the production of economists.

The shifts in the subject areas of dissertations over time indicate that the areas

*Associate professor, University of Iowa and Coe College; and graduate student at Northwestern University, respectively. We wish to thank the Baker Foundation for partial support and the Coe College Computer Center for assistance. Calvin Siebert and Richard Doyle made significant improvements on earlier drafts of the paper.

[1] The percentages after 1940 are for conferred degrees, whereas the percentages for 1904–39 include all candidates. In addition, Froman's data for 1928 included a significant duplication which has been corrected. The data for the period 1904–39 are from Froman and data from 1940 were directly compiled for all tables.

[2] The 1904–39 listing excludes Johns Hopkins, Northwestern, Princeton, Iowa, Catholic, Brookings, and Radcliffe, which were in the top twenty during that period. Northwestern, Princeton, and Iowa were also in the top twenty during 1940–59; and UCLA ranked sixteenth for 1960–74.

[3] During 1904–28, the top three schools produced 50.8 percent of the candidates. This figure was only 15.5 percent for the most recent period analyzed, 1970–74.

TABLE 1—THE TOP TWENTY PH.D. SCHOOLS
(Shown in Percent)

	1904-39	1940-59	1960-69	1970-74	1940-74
1. Harvard	8.7	12.7	6.3	5.3	8.0
2. Berkeley	2.4	4.5	6.5	6.2	5.8
3. Columbia	20.8	9.4	5.6	1.6	5.1
4. Wisconsin	9.3	8.5	2.6	3.7	4.7
5. Chicago	13.4	7.6	3.2	2.8	4.4
6. N.Y.U.	1.0	5.0	2.7	2.0	3.2
7. Indiana	.2	4.2	3.8	1.3	2.9
8. Michigan	2.0	2.3	3.6	2.4	2.9
9. Minnesota	3.4	3.5	2.9	1.5	2.7
10. Pennsylvania	6.3	3.0	3.0	1.9	2.7
11. Iowa State	a	3.2	2.2	2.5	2.6
12. M.I.T.	a	2.0	3.4	1.5	2.5
13. Cornell	4.1	2.0	2.4	2.6	2.4
14. Michigan State	a	a	2.4	4.0	2.2
15. Stanford	1.5	.5	3.3	2.2	2.1
16. Ohio State	2.5	3.0	2.0	1.4	2.1
17. Texas	.3	3.2	1.8	1.3	2.1
18. Purdue	a	.6	2.1	3.5	2.0
19. Yale	2.6	1.5	2.5	1.8	2.0
20. Illinois	3.1	1.9	1.3	2.4	1.8
Top 20	91.7	82.3	63.6	51.9	64.2
Top 40	99.2	90.5	84.4	67.8	84.2
Others	.8	9.6	15.6	32.2	15.8
Total Number of Degrees Granted	5620.0	3918.0	5339.0	3535.0	12792.0

[a]Denotes the school is not in top forty during the era.

of economic history and history of thought have declined substantially; also declining is the area of welfare and consumer economics which was classified as social problems in the earlier periods. This can be accounted for by the increased specialization in subject areas and the deemphasis of dissertations in institutional economics. This shift is not nearly so predominant in the professional literature over this time period, as noted by Martin Bronfenbrenner's study, but the direction of change is consistent in both trends.

Table 2 shows the distribution of theses by topic for three distinct periods. The subject categories in this table are a synthesis of the various general topical breakdowns used in the *AER* dissertation lists from 1904 to 1974.[4] The categories used by the *AER* to classify theses by subject have fluctuated considerably in some instances. Careful judgment was of the essence in combining categories so as to be certain that these were consistently classified, thereby allowing an intertemporal comparison of topical emphasis.

Table 2 indicates that the intensity of investigation of different economic topics varies tremendously over time. From 1904 to the present, business administration, international economics, and economic growth and development have grown in stature as dissertation topics. Concurrently, the institutional areas of economic history, the history of economic thought, and welfare and consumer economics have declined in importance. These trends coincide with the tendency to the preparation of dissertations employing more advanced empirical techniques.[5] The trends in choice of dissertation

[4]Categories were synthesized using the *Index of Economic Articles* classification system.

[5]Gabriel provides an interesting breakdown of dissertations by technique and method between 1904 and

TABLE 2—DOCTORAL DISSERTATIONS BY SUBJECT AREA
(Shown in Percent)

	1904-28	1929-40	1970 74	1940-74
Economic Theory	5.9	5.6	6.0	6.0
Economic History and History of Thought	13.2	6.9	3.0	4.0
Agriculture	9.1	12.9	10.0	13.0
Industrial Organization	8.9	8.8	9.0	10.0
International Economics	4.3	5.6	9.0	8.0
Business Administration	8.1	10.9	9.0	14.0
Economic Systems	3.3	3.9	1.0	1.0
Labor	12.7	8.9	9.0	10.0
Monetary Theory and Institutions	6.4	12.5	8.0	7.0
Fiscal Theory and Public Finance	8.4	9.0	4.0	6.0
Population and Urban-Regional	3.8	2.4	8.0	3.0
Welfare/Consumer Economics	13.8	10.4	5.0	3.0
Statistics and Econometrics	1.9	2.2	5.0	3.0
Economic Growth and Development	–	·	13.0	11.0
Total	99.8	100.0	99.0	99.0

topics are also generally consistent with the movement of journal article subject emphasis.[6]

Table 3 provides a breakdown of the dissertation production of the leading twenty schools from 1940-74 according to the subject area distribution. The percentages in the table show each school's share of the total dissertations in each respective subject area. The bottom row shows each school's overall contribution of dissertations during 1940-74.

The instances in which a school contributed a portion of the theses in a given subject area which was at least twice as great as its overall percentage of dissertations indicates some degree of specialization in the given topic by the respective institution. This criterion suggests that M.I.T. and Yale concentrate their efforts in the general theory area. Illinois is the relative leader in the history of economic thought. Economic history is dominated by Chicago, Columbia, and Yale. Yale alone specializes in economic systems. Purdue and Michigan produce a disproportionate number of econometrics dissertations. Pennsylvania and Illinois dominate the field of social accounting, and business administration studies are prevalent at Indiana, Michigan State, and Stanford. Stanford, Yale, and Illinois prevail in industrial organization while Iowa State, Minnesota, Purdue, Cornell, and Ohio State demonstrate superiority in agricultural economics. Not surprisingly, these five schools are all land-grant institutions.[7] Finally, the areas of labor economics and economic welfare are both dominated by Wisconsin. The other areas do not have any one top twenty school which meets the criterion of producing twice as many dissertations in the respective categories as its overall output of Ph.D.s. Princeton, which ranks twenty-first for the period, does show this relative concentration in population studies. U.S.C., twenty-fourth for the period, exhibits a similar strength in economic systems. And Duke, which ranks twenty-sixth overall for 1940–74, shows relative specialization in the history of economic

1974. This study also has more complete data from which the summary tables in this paper were developed.

[6]Note Bronfenbrenner's study on journal topics over time.

[7]Table 3 shows that Harvard produced a surprising number of dissertations in agricultural and natural resources areas; Harvard produced one-fifth of the dissertations in this field during the 1940's. This would seem to discount Orville Freeman's claim that the only reason he was Secretary of Agriculture under President Kennedy was because Harvard didn't produce agricultural economists or have a school of agriculture.

TABLE 3—SUBJECT AREA OF DISSERTATION FOR TOP INSTITUTIONS, 1940–74
(Shown in Percent)

	Harvard	Berkeley	Columbia	Wisconsin	Chicago	N.Y.U.	Indiana	Michigan	Minnesota	Penn	Iowa State	M.I.T.	Cornell	Mich. State	Stanford	Ohio State	Texas	Purdue	Yale	Illinois	Others
General Theory	10 0	5 6	4 0	1 6	4 3	1 6	1 0	2 0	2 7	2 7	.7	7.3	1.0	.7	3.7	.7	2.0	2.7	4.3	2.0	39.4
History of Thought	3 8	4.8	7.6	5.7	4.8	3 8	1.9	9	.9	1.9	.9	.9	1.9	.6	9	9	2 9	.9	.6	4.8	48.6
Ec History	9 3	4 1	10 3	4 1	15.5	3 1	2 1	4	4	2 1	2 1	1.0	1.0	.4	1.0	1.0	3.1	3 1	4.1	1.1	31 7
Ec Systems	15 5	5.2	4 1	1 1	6.2	1.0	1.1	4	.3	4	3.1	1 1	1 1	4	1.0	1.1	4.1	1 1	4.1	1.1	46.5
Growth/Devel	10 0	8.2	4.1	4.0	2.1	3.0	1.9	2.0	3.2	.9	2.0	3.1	2.9	3.9	1.9	3.1	2.8	1.1	3.2	1.8	31.4
Fluct /Forecasting	15 7	7.2	10 0	3 4	6 6	1 0	3 2	3 8	2.0	2 9	.4	2 9	.8	.9	.9	1 0	2.0	1 0	3.1	.9	30 2
Econometrics	6 1	6 1	2 0	4 1	4.0	4.0	1 9	6.0	3.1	2.1	2 2	3.0	2 1	3 1	.8	1 0	2 1	6 1	2 0	2.1	26.1
Social Acct	6.7	7 7	4 8	6.6	1 9	3	4	3.8	2.9	8 7	3 8	1 9	.7	3.6	1.8	.8	1.9	1 9	9	3.8	35.1
Monetary Theory	6 8	4 2	6.0	2 8	7.4	4 0	3 9	2.9	2.8	3 0	3.1	3 0	7	.9	2.1	2.0	3 3	1 8	2.1	2.0	35 2
Fiscal Theory/ Pub Fin	6.9	3.1	7 2	8 0	3.0	5.2	2.9	3 0	1.9	.9	.9	2.0	.9	.8	.8	.9	2 0	.9	1.9	3.1	43.5
International	9 2	4 1	6 1	4 0	6 2	2 9	8	2 9	1 1	2 0	8	4 1	2 0	3 1	3 1	1 8	1 7	8	3 1	1 9	38 2
Bus Adm	6 9	4 9	5.0	2 0	2 9	6 1	6 0	4.9	.8	4 8	1 1	1 0	1 0	4 9	4 9	3 9	3.9	1 9	1.0	1.0	32.0
Ind Org.	12 7	9 8	1 9	3	5 8	3.9	4.0	3.8	.4	4.0	.2	1 9	.3	4	7.8	2	3.9	.4	5.8	3.9	32.2
Agriculture	7 2	8 4	2 1	6.2	4 1	.6	7	3	10 3	1.0	12 4	6	5 2	3 1	2 1	5.2	8	6 2	.2	2.1	21.2
Labor	6 3	6 2	5.4	9 8	5 3	3 4	1 5	2 4	1 5	3 4	.4	4 4	3 9	4	7	1.5	1 5	4	5	1.4	39.4
Population	5.9	4 8	9 1	2 8	7 2	5 1	8	3 8	1.1	3.0	2	2 0	3 9	1 0	1.0	3	2 1	4	9	1.0	47.3
Urban/Regional	3 9	3 8	2 0	2 1	3 1	.8	.8	2 9	1 0	2 0	2 9	1 9	2 0	8	3.0	1 0	.3	.9	2 0	3 1	59 7
Welfare/Consumer	8 1	6.0	5 1	9 1	3 9	2.8	2.7	2.9	2 8	2.7	3 0	1.9	2 1	1.9	1.8	2.1	2.0	1.9	2.0	2.1	33 1
Percent of Total	8.0	5.8	5.1	4 7	4 4	3.2	2.9	2 9	2 7	2.7	2.6	2 5	2 4	2.2	2 2	2.1	2.1	2.0	2.0	1.8	35.8

thought, economic systems and population studies.

It was previously noted that there has been a dispersion of the production of Ph.D.s in economics with respect to graduate institution attended. Simultaneously, the concentration particular schools possessed in certain subject areas prior to 1940 has also been lessened. Although there is demonstrable specialization by particular schools, no one school dominates a field to the extent common of dissertation production prior to 1940. The data for successive decades show a progressive tendency away from the monopolization of particular fields by a few schools and continuing growth in dispersion.

Table 4 shows the top twenty schools for granting bachelor's degrees to those that received the Ph.D. There has been some dispersion in undergraduate origin of Ph.D.s, but only in the last thirty-five years. Undergraduate schools not in the top twenty accounted for 65.3 percent of the 1904 28 doctorates, and only 59.3 percent of the 1929–40 Ph.D.s. During the past thirty-five years, however, the small decrease in the dispersion of undergraduate study between 1904 and 1940 was strongly reversed. Seventy-two percent of the economists who received their Ph.D.s during 1940-74 studied as undergraduates at schools other than those in the top twenty.

The percentage of Ph.D.s who received their undergraduate training at foreign schools increased from 6 percent in the 1904-40 era to 13 percent between 1940 and 1974. Foreign-born economists who received their doctorates in the United States have not been discriminated against by the more prestigious institutions. The top ten institutions granting doctorates to the foreign born are all listed in the top twenty Ph.D. schools for 1970-74. The forty-three schools in the "chairmen's group" produce approximately 75 percent of the doctorates and 83 percent of the foreign-born economists. The shift in the top bachelors' schools has been from the liberal arts schools prior to 1940 to state universities since 1940. From 1904 to 1928, fifteen of the top forty were liberal arts institutions, but only Oberlin, Swarthmore, Amherst, and Williams remain in the top forty for 1940–74. However, when undergraduate school's contribution of Ph.D.s was standardized for enrollment differences, thirty of the top forty schools were liberal arts institutions. Berkeley and Cornell were the only state universities which remained in the top forty

TABLE 4—TOP TWENTY SCHOOLS GRANTING BACHELOR'S DEGREES TO ECONOMICS DOCTORATE RECIPIENTS, 1940–74

	Number	Percent
Berkeley	170	2.4
Harvard	170	2.4
C.C.N.Y.	138	2.0
Illinois	113	1.6
Wisconsin	110	1.6
Michigan	105	1.5
Minnesota	104	1.5
Cornell	99	1.4
Columbia	91	1.3
N.Y.U.	86	1.2
Chicago	83	1.2
Texas	82	1.2
Pennsylvania	76	1.1
U.C.L.A.	73	1.0
Iowa State	73	1.0
Ohio State	67	1.0
Oberlin	66	.9
Brooklyn College	66	.9
Yale	65	.9
U. of Washington	64	.9
Others	5069	72.4

Note: This total includes only those who listed their school in the various editions of the *Handbook* of this *Review* or were in the *American Men and Women in Science: Economics, 1974*.

after the adjustment for enrollment differences, and the other eight schools were private universities. The undergraduate training has been divided equally between private and public institutions.

Table 5 presents the distribution of Ph.D.s during 1940 74 by sex.[8] Froman first collected data on the sex of 1929–40 degree recipients. During those years, an average of 8.0 percent of the Ph.D.s were women, and the percentage ranged annually from 4.7 to 12.9 percent. The proportion of doctorates in economics who have received their degrees since 1940 and are female has declined substantially. An average of only 5.2 percent of the 1940–74

TABLE 5—WOMEN DOCTORATES IN ECONOMICS, 1929–74

	Total	Percent
1929-40[a]	225	8.0
1941 45	27	5.2
1946-50	49	6.9
1951 55	76	5.1
1956-60	70	4.5
1961-65	88	4.0
1966-70	161	4.4
1971-74	163	6.1
1940-74	636	5.0

[a]The 1929-40 numbers are from Froman (1942, p. 825).

Ph.D. recipients were women. The percentage ranged from an annual high of 11.2 percent in 1947 to a low of 2.5 percent in 1940. There may be a trend to greater participation by women, however. During the 1960's only 4.0 percent of the Ph.D.s were females, but this statistic rose to 6.0 percent for 1970–74. The undergraduate training of women economists has been very concentrated as one-fourth of the degrees have been from the seven sister institutions; however, their graduate training has been distributed among the top-rated schools in a proportional manner.

To summarize: the most prolific producer of economists in the 1970's has been the University of California at Berkeley. The most intensively investigated subject areas of economic doctoral dissertations are currently economic growth, development, and planning, and the economics of natural resources and agriculture. The great majority of these recent dissertation authors received their undergraduate preparation at private and state universities which are not among the major graduate schools. And over 90 percent of these economists who received their Ph.D.s in the last twenty-five years are male.

[8]These female-grouping data were obtained by classifying by feminine first names; however, we are aware of the "Sally Frankel" phenomenon that W. Lee Hansen and Burton Weisbrod discovered after Mr. Frankel was placed on their elite list of women who had published in economic journals.

REFERENCES

M. Bronfenbrenner, "Trends, Cycles, and Fads in Economic Writing," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 538–52.

L. Froman, "Graduate Students in Economics, 1904 to 1928," *Amer. Econ. Rev.*, June 1930, *20*, 235–47.

——, "Graduate Students in Economics, 1904–40," *Amer. Econ. Rev.*, Dec. 1942, *32*, 817-26.

——, "Graduate Students in Economics," *Amer. Econ. Rev.*, Sept. 1952, *42*, 602–08.

D. B. Gabriel, "An Empirical Treatment of 20th Century American Economic Thought," unpublished thesis, Coe College 1976.

W. E. Spellman and G. Holland, "A Note on the Status of Women in Economics," *J. Econ. Educ.*, Spring 1976, *7*, 124–25.

B. Weisbrod and W. L. Hansen, "Towards a General Theory of Awards, or Do Economists Need a Hall of Fame," *J. Polit. Econ.*, Mar./Apr. 1972, *80*, 422–31.

American Economic Review, Handbook, various issues.

American Men and Women in Science: Economics, New York; London 1974.

"Report of the Committee on the Status of Women in the Economics Profession," *Amer. Econ. Rev. Proc.*, May 1973–76, *63–66*.

"Thirty-Seventh (through seventy-first) List of Doctoral Dissertations in Political Economy . . . in American Universities and Colleges," *Amer. Econ. Rev.*, 1940-74, *30–64*.

The Role of Money in a Simple Growth Model: Note

By WILLIAM B. MARXSEN*

In a recent article in this *Review*, David Levhari and Don Patinkin (hereafter noted L-P) develop an equilibrium growth model for a simple economy in which money is treated as a productive factor, entering an aggregate production function. In their policy section, they are unable to determine the effects of an increase in the exogenously determined rate of inflation on the equilibrium capital and real-balance intensities. They are also unable to establish whether or not steady-state equilibrium is stable. The purpose of this note is to extend their dynamic analysis and to show that 1) the effects of a change in the rate of inflation are more or less predictable, and 2) steady-state equilibrium can be expected to be stable.

The L-P model can be briefly summarized. Assume a growing neoclassical economy where $Y = G(K, M/P, N)$, and where Y, K, M/P, and N represent real output, the stock of physical capital, the real money stock, and the labor force, respectively. Assume further that G is linear homogeneous and twice continuously differentiable. The function can therefore be written $y = g(k,m)$, where y, k, and m are Y/N, K/N, and M/PN, respectively. Assume g is well-behaved such that $g_i \geqq 0$, $g_{ii} < 0$, and $g_{ij} = g_{ji} > 0$. Let the labor force grow at some exogenously determined exponential rate n.

Money is costlessly produced by the government and injected into the economy via transfer payments to the public. In order to avoid stability problems noted by Miguel Sidrauski, it is assumed that the rate of nominal expansion is altered by the government in order to maintain a constant target rate of inflation.[1] The rate of nominal expansion is $u = DM/M$, where D denotes the time derivative of the variable which follows it. The rate of inflation is $p = DP/P$. It follows that $D(M/P) = (u - p)M/P$. From this and the labor force growth assumption, it follows that the rate of growth of the real per capita money stock is

$$(1) \qquad Dm/m = u - p - n$$

The demand for money derives from the profit-maximizing behavior of firms where it is assumed that the return to capital and real balances will be equalized. The return to capital equals its marginal physical product, $g_k > 0$. The return to money includes, however, not only its marginal physical product $g_m \geqq 0$, but the rate of deflationary appreciation in its value $-p$ as well. Consequently, the L-P equation for asset equilibrium is

$$(2) \qquad g_k(k,m) = g_m(k,m) - p$$

Equation (2) can be solved implicitly for the demand for real balances, m_d, in terms of k and p, where

$$m_d = L(k,p)$$

$$\frac{\partial m}{\partial k} = L_k = \frac{g_{kk} - g_{mk}}{g_{mm} - g_{km}} > 0$$

$$\frac{\partial m}{\partial p} = L_p = \frac{1}{g_{mm} - g_{km}} < 0$$

*Assistant professor of economics, University of Montevallo.

[1] In his comment on the L-P model, Jon Harkness replaces this assumption with government maintenance of a constant rate of nominal expansion of the money supply, where the rate of inflation is free to vary during steady-state disequilibrium. L-P note this in their "Reply" and are puzzled by the comparative dynamic implications of the Harkness conclusions. They ignore, however, Sidrauski's demonstration that steady-state equilibrium in an economy characterized by static money-market equilibrium, perfect inflationary foresight, and the Harkness nominal expansion assumption is *at best* a saddlepoint. Consequently Harkness' invocation of the Correspondence Principle is not legitimate and his conclusions are erroneous.

Furthermore, money-market equilibrium is assumed so that the amount of real balances demanded equals the money supply.

The value of the capital stock is determined by consumption-saving behavior of the public. For simplicity L-P assume that a fixed proportion s of disposable income is saved, where disposable income equals real output plus the rate of real monetary transfer to the public, $(u - p)M/P$. Saving, however, is divided between the accumulation of physical capital and of real balances. Thus,

$$s[Y + (u - p)M/P] = DK + (u - p)M/P \quad (3)$$

This is L-P's equation (42), (1968, p. 738). By making use of the fact that $DK/N = Dk + nk$, equation (3) may be restated in per capita terms:

$$Dk = sy - (1 - s)(u - p)m - nk \quad (4)$$

In steady state, real balances and the capital stock will grow at the same rate as the labor force, so that $Dm = Dk = 0$. From equation (1) it will then be true that the rate of nominal expansion will exceed the rate of inflation by n, so that the monetary component of disposable income and saving will equal nm, regardless of the rate of inflation. General steady-state equilibrium will obtain when $Dk = 0$ and the money market is in equilibrium, or when

$$g(k, m) = \frac{nk}{s} + (1 - s)\frac{nm}{s} \quad (5)$$

and

$$m = L(k, p) \quad (6)$$

At this point in their analysis, L-P attempt to determine the effects of a change in p on the steady-state values of k and m by solving equations (5) and (6) for dk/dp and dm/dp. They obtain

$$\frac{dk}{dp} = \frac{sg_m + (s - 1)n}{-\Delta} \quad (7)$$

$$\frac{dm}{dp} = \frac{sg_k - n}{\Delta} \quad (8)$$

where

$$\Delta = \frac{sg_k - n + L_k[sg_m - (1 - s)n]}{L_p} \quad (9)$$

Equations (7)–(9) are L-P's equations (53)–(55), (1968, pp. 740–41). Levhari and Patinkin suggest that the earlier assumptions concerning production, the demand for money, and consumption are not sufficient to determine the signs of $sg_k - n$ and $sg_m + (s - 1)n$, and hence of the derivatives (7) and (8).

The difficulty lies in the indeterminacy of the slope of the line showing the locus of all possible steady-state combinations of k and m, those satisfying equation (5). The steady-state line, however, can be constructed by first plotting various components of equation (5). The left-hand side of the equation is shown in Figure 1. The usual isoquant map for the production function is plotted where an arbitrary unit of measurement for output has been chosen. Each subsequent unit of output requires progressively larger increases in k and m (services) due to the fixity of labor, and hence, the isoquants get farther apart as k and m are increased. In Figure 2 isoquant lines for $nk/s + (1 - s)\cdot(nm/s)$ are plotted (in the same units as

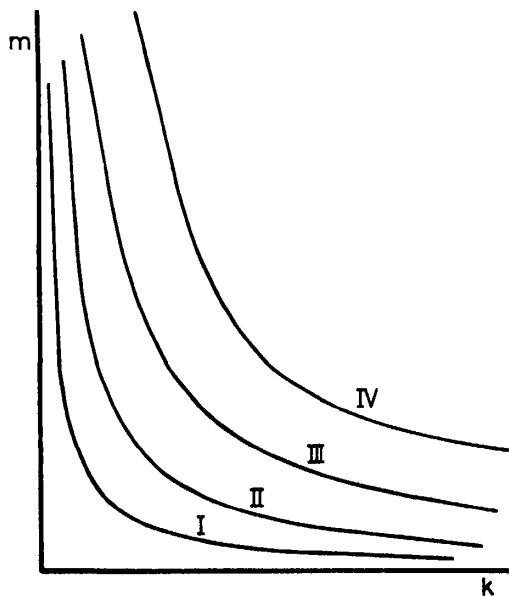



FIGURE 1

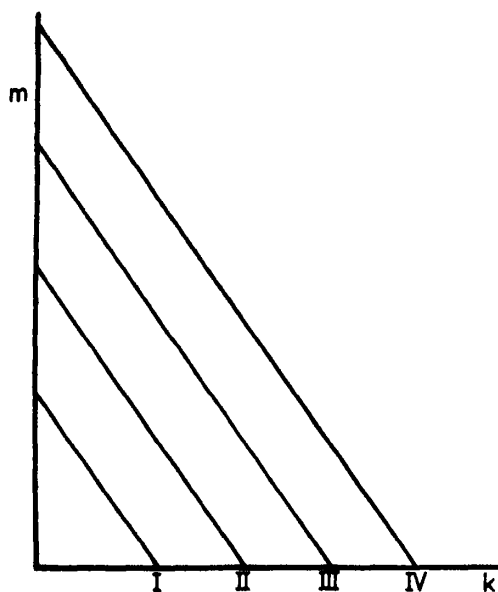



FIGURE 2

output) for the same values as the output isoquants. Unlike the production isoquants, these are equidistant, since doubling k and m doubles any linear combination of the two.

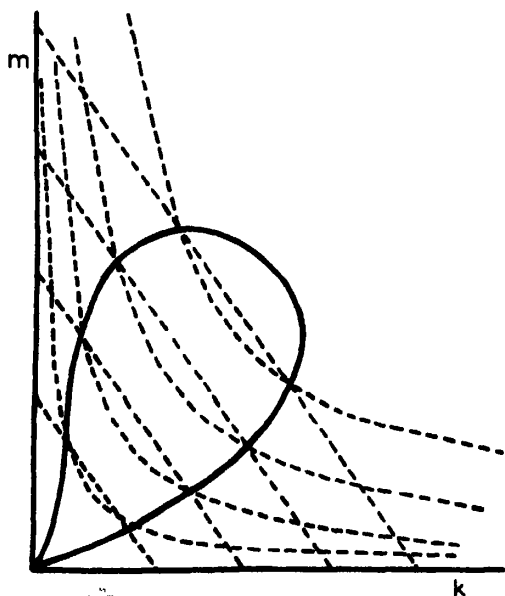



FIGURE 3

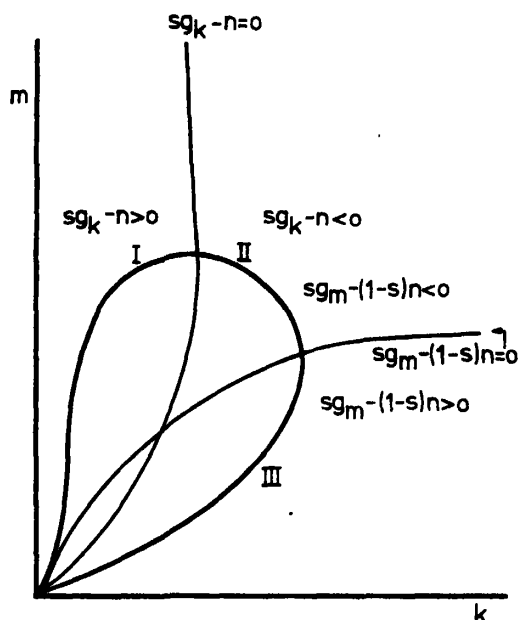



FIGURE 4

Figures 1 and 2 have been superimposed in Figure 3 and the locus of the intersections of the equivalued isoquants is shown. The heavy curve shows all the combinations of k and m satisfying equation (5) and is L-P's *cc* line. In Figure 4, starting at the origin and proceeding clockwise along the *cc* line, it is clear that there are three distinct regions. In Region I, capital is scarce relative to both money and labor, so that $sg_k - n > 0$,[2] but money's abundance relative to k makes g_m low, such that $sg_m - (1 - s)n < 0$; in Region II, both capital and real balances are abundant, so that both $sg_k - n$ and $sg_m - (1 - s)n$ are less than zero; and finally, in Region III, K and N are abundant relative to M/P, so that $sg_k - n < 0$ and $sg_m - (1 - s)n > 0$. These values are in accord with the slope of the steady-state line, as may be verified by dif-

[2]This contradicts Harkness, who claims that $sg_k - n$ "is unequivocally negative" (p. 178). He correctly demonstrates that $sg_k - n = -s(w + um)/(k + m)$, where w equals the wage rate, but ignores the possibility that u might be negative and substantial enough to make $w + um$ negative, as well, and therefore $sg_k - n$ might be positive.

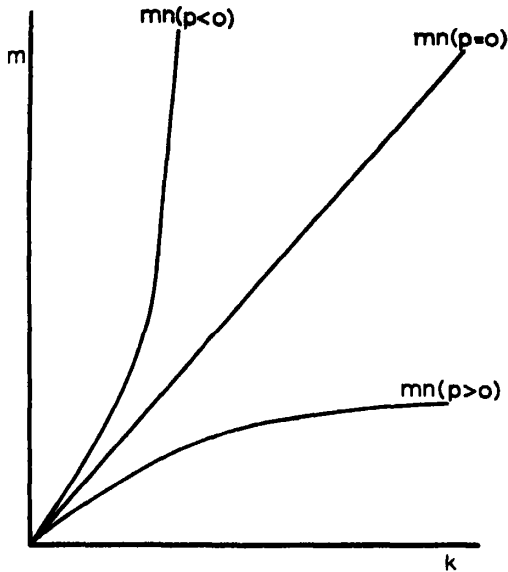



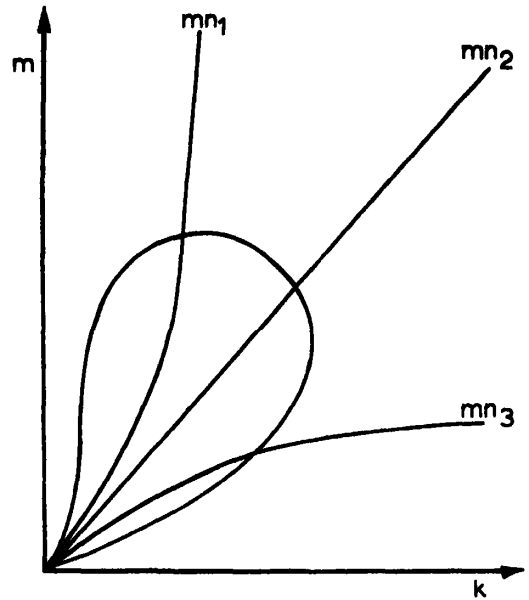
FIGURE 5

FIGURE 6

ferentiation of equation (5) with respect to k:

$$\left(\frac{dm}{dk}\right)_{cc} = -\frac{sg_k - n}{sg_m - (1 - s)n} \tag{10}$$

The position of the cc line is determined entirely by the characteristics of the production function, the thrift of the public, and the rate of growth of the labor force. Changes in p by the monetary authority have no effect on its position. Changes in p affect the composition of the public's portfolio of capital and money, and therefore shift the locus of $k - m$ combinations which satisfy money-market equilibrium equation (6). Recall from earlier analysis that $\partial m/\partial k = L_k > 0$ and $\partial m/\partial p = L_p < 0$. The mn curves which follow from these derivatives are shown in Figure 5. Note that for $p < 0$, there is a maximum k at which mn becomes vertical. At this k, $g_k \rightarrow -p$ and $g_m \rightarrow 0$. For $p > 0$, there is a similar m value which is approached asymptotically, where $g_k \rightarrow 0$ and $g_m \rightarrow p$. These properties follow from the restrictions imposed by equation (2).

General steady-state equilibrium occurs at the intersection of cc and mn. Such equilibria are shown in Figure 6 for three different rates of inflation. As the rate of inflation is increased, mn rotates from mn_1 through mn_3 where such increases initially raise both equilibrium capital and real-balance intensity, then raise capital while lowering real-balance intensity, and then ultimately lower both intensities. These results are of limited usefulness, since they fail to indicate the range of inflation rates for which equilibrium will fall in the various regions of the cc curve and hence for which increases in the rate of inflation will have Tobinesque effects on capital and real-balance intensities.

Greater precision concerning neutrality of monetary policy can be achieved by recalling equation (10) for the slope of the cc line and solving for the relationship between the rate of inflation and the values of $sg_k - n$ and $sg_m - (1 - s)n$. From the assumption that the production function is linear homogeneous, it follows that

$$g(k,m) = w + g_k k + g_m m \tag{11}$$

where w equals the wage rate. The portfolio-balance condition (equation (2)) may be substituted into equation (11) to obtain

$$g(k,m) = w - pk + g_m(k + m)$$

The right-hand side of the above equation may then be substituted into the steady-state condition (equation (5)) to obtain

$$w - pk + g_m(k + m) = \frac{n}{s} k + \frac{n}{s}(1 - s)m$$

Recalling that in steady state $p = u - n$, and manipulating the above equation, yields

$$(12) \quad sg_m - (1 - s)n = -\frac{s(w - uk)}{k + m}$$

Following a similar procedure yields the value of $sg_k - n$:

$$(13) \quad sg_k - n = -\frac{s(w + um)}{k + m}$$

Equations (12) and (13) show the relationship between u (and therefore p) and $sg_k - n$ and $sg_m - (1 - s)n$. For low rates of nominal expansion or contraction, w will exceed both uk and um and therefore both $sg_m - (1 - s)n$ and $sg_k - n$ will be negative. Therefore, for rates of nominal expansion in the neighborhood of $u = 0$, general steady-state equilibrium will occur in Region II of the cc curve, where an increase in u and therefore p will tend to raise steady-state k and lower steady-state m, similar to the standard Tobinesque money-growth model result. For very high rates of nominal *contraction*, $sg_k - n > 0$, and steady-state equilibrium will occur in Region I. Here a reduction in the rate of deflation (increase in u) will raise both m and k, enhancing steady-state output and consumption. Finally, for very high rates of *inflation*, $sg_m - (1 - s)n > 0$, and equilibrium will occur in Region III. An increase in the rate of inflation will reduce both m and k.

The last two cases indicate the limitations imposed on the monetary authority when money is a productive asset and changes in the rate of inflation not only alter the saving behavior of the public but affect the allocation of productive resources as well. Policies of high inflation or deflation so distort the public's holdings of money and capital that steady-state output and consumption are lower than with moderate policies. Levhari and Patinkin demonstrate that the consumption-maximizing rate of nominal expansion of the money supply is zero (see their equations (69), (70), 1968, pp. 747–48), and that the optimum monetary policy is one of complete inactivity. In such a case, mn intersects cc where its slope is minus unity and steady-state wealth $(k + m)$ is maximized (this follows from equations (10), (12), and (13)).

So long as mn intersects cc from inside as shown in Figure 6, steady-state equilibrium will be locally stable. Since p is held constant by the monetary authority and m is confined to equilibrium values, the dynamic behavior of m is linearly dependent on the behavior of k. Likewise u, the only other dynamic variable, is similarly dependent on m by virtue of equation (1). Consequently, the stability test reduces to one dimension. Differentiation of equation (6) with respect to time (assuming p constant) yields

$$(14) \quad Dm = L_k Dk$$

Substituting equation (14) into equation (1), substituting the resulting equation into equation (4), and then differentiating with respect to k, yields

$$(15) \quad \frac{d(Dk)}{dk} = \frac{sg_k - n + L_k[sg_m - (1 - s)n]}{1 + (1 - s)L_k}$$

The denominator of the term on the right-hand side of equation (15) is unequivocally positive, and consequently equilibrium is stable if and only if the numerator is negative. In Region II of the cc line, where equilibrium will be found for moderate rates of inflation or deflation, both $sg_k - n$ and $sg_m - (1 - s)n$ are negative, guaranteeing that the numerator will be negative (since $L_k > 0$). Hence, in this region stability is guaranteed.

But in Regions I and III the signs of $sg_k - n$ and $sg_m - (1 - s)n$ differ, so that stability is no longer guaranteed. The numerator may be rewritten as

$$(16) \quad \text{Num} = -(sg_m - (1 - s)n) \cdot \left(\frac{-sg_k - n}{sg_m - (1 - s)n} - L_k\right)$$

From equations (6) and (10), (16) can be rewritten as

$$(17)\quad \text{Num} = -(sg_m - (1 - s)n) \cdot (\text{slope } cc - \text{slope } mn)$$

In Region I, $sg_m - (1 - s)n < 0$ and if mn intersects cc from within, then slope $mn >$ slope $cc > 0$. Consequently, the numerator is negative and stability is guaranteed. In Region III, $sg_m - (1 - s)n > 0$. Here, if mn intersects cc from within, then slope $cc >$ slope $mn > 0$. Again, the numerator will be negative and equilibrium will be stable.

The graphical analysis has shown that for moderate inflation and deflation rates an increase in the rate of inflation will raise equilibrium capital intensity and lower that of real balances. This is in accord with the usual Tobinesque result and suggests that the Tobinesque effect of the altered transfers on disposable income is larger than the portfolio effects on production of such a change. For high rates of inflation and deflation, however, the production effects dominate so that the changes in equilibrium capital and money intensities differ substantially from the Tobinesque predictions. Finally, the graphical approach was useful in demonstrating that with the L-P expansion mechanism, steady-state equilibrium is stable, regardless of the rate of inflation chosen by the monetary authority.

REFERENCES

J. Harkness, "The Role of Money in a Simple Growth Model: Comment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 177–79.

D. Levhari and D. Patinkin, "The Role of Money in a Simple Growth Model," *Amer. Econ. Rev.*, Sept. 1968, *58*, 713-53.

———, "The Role of Money in a Simple Growth Model: Reply," *Amer. Econ. Rev.*, Mar. 1972, *62*, 185.

M. Sidrauski, "Inflation and Economic Growth," *J. Polit. Econ.*, Dec. 1967, *75*, 796–810.

A Mean-Standard Deviation Exposition of the Theory of the Firm under Uncertainty: A Pedagogical Note

By GABRIEL A. HAWAWINI*

In his paper in this *Review*, Agnar Sandmo derived a set of major conclusions indicating that a competitive firm behaves differently under uncertainty than in a world of certainty. Hayne Leland extended the results to noncompetitive market structures. The purpose of this paper is to show that firms' behavior under uncertainty can be easily derived using a geometric approach based on the mean-standard deviation framework introduced by Harry Markowitz (1952, 1959) and extended by James Tobin.

Section I discusses briefly the major conclusions related to the firm's behavior under uncertainty and gives a general description of the approach followed in this paper. Section II introduces the firm's attitude toward risk. Section III describes a model of profit maximization under conditions of risk, using a mean-standard deviation-of-profit framework. In Section IV, the model is applied to derive geometrically the major conclusions of the theory of the firm under uncertainty stated in Section I.

I. The Theory of the Firm under Uncertainty

A. *General*

Traditional microeconomic theory assumes that under certainty and regardless of the market structure, a firm's objective is to maximize its profit for the given constraints. The optimal output is obtained at the point at which the firm's marginal cost equals its marginal revenue. However, if uncertainty prevails, there is no reason to believe, a priori, that this maximization principle will hold.

Both Sandmo and Leland have used the assumption that faced with uncertainty, the firm will maximize the expected value of its utility of profit. The introduction of a nonlinear utility function[1] permits the incorporation of the firm's attitude toward risk into the decision-making process. Sandmo assumes a subjective probability distribution of prices with the level of output and cost function known in advance, that is, under the firm's control. Consequently, since the firm is unable to influence the price distribution, it is considered a price taker, and Sandmo's model is valid only under condition of perfect competition. Leland's model is more general. It assumes a random demand function that allows us to handle noncompetitive structure, where a firm can fix the level of output and/or the price. Sandmo's conclusions are shown to be a special case of Leland's model when the market is competitive.

The major conclusions that follow from the theory of the firm operating under uncertainty are: (i) if a firm is risk averse[2] its optimal output is smaller than the certainty output; (ii) if a firm displays decreasing absolute risk aversion, its optimal output varies inversely with its fixed costs; (iii) if a competitive firm displays decreasing absolute risk aversion, it has an upward-sloping supply curve; (iv) if a firm is risk averse, an equilibrium exists, even under constant or decreasing marginal costs; (v) if a firm is

*Assistant professor, New York University. I would like to thank Robert Schwartz, New York University; Roger Mesznik, Baruch College; and an anonymous referee for useful comments.

[1]For linear utility functions, the firm will be maximizing expected profit which implies that the firm is indifferent to the magnitude of risk that is associated with the production under uncertainty: the firm is risk neutral.

[2]Risk aversion is defined in Section IIA.

risk averse, equilibrium requires the existence of positive profit.

In addition to the above conclusions derived by Sandmo and Leland, I will show that: (vi) if a firm displays decreasing absolute risk aversion, its optimal output varies inversely with its perceived level of risk; (vii) if a firm displays a decreasing absolute risk aversion, its optimal output varies inversely with its variable costs; (viii) under uncertainty the competitive firm will produce a higher output than the noncompetitive firm selling at the same price.

B. *Risk, Expected Profit, and Equilibrium: A Description of the Mean-Standard Deviation Approach*

In this paper I attempt to picture the firm's optimum output under price uncertainty in a mean-standard deviation-of-profit plane, with the standard deviation of profit considered as a proxy for risk. On this plane the firm's indifference map is drawn, a geometrical representation of its attitude toward risk. In Section III, I derive the relationship between expected profit and the standard deviation of profit (risk), which is called "the profit-opportunity locus," and drawn on the same plane. Equilibrium is found at the point where the profit-opportunity locus and the firm's highest indifference curve are tangent. From this equilibrium point it is shown that the corresponding level of optimum output under uncertainty can be easily derived.

II. The Firm's Attitude Toward Risk

A. *Definitions*

A firm is assumed to have a utility function of profit that displays positive marginal utility of profit. The firm's attitude toward risk is indicated by the change in its marginal utility when profit varies. A firm is said to be risk averse if its marginal utility of profit decreases with increasing profit, risk neutral if its marginal utility of profit is constant, and a risk seeker if its marginal utility of profit increases with increasing profit.[3] In this paper, firms are assumed to display risk aversion.

A risk-averse firm is said to display decreasing absolute risk aversion if its risk aversion decreases with increasing profit, constant absolute risk aversion if its risk aversion remains constant when profit changes, and increasing absolute risk aversion when its risk aversion increases with increasing profit.[4]

B. *Indifference Curves in the Mean-Standard Deviation-of-Profit Plane*

The expected utility of profit can be written as a function of the first two moments of the distribution of profit.[5] An indifference curve in the mean-standard deviation-of-profit plane is then represented by the locus of points for which the expected utility of profit remains constant. An indifference map is generated by varying the constant value of the expected utility of profit. In the mean-standard deviation-of-profit plane, the risk-averse firm has indifference curves with positive marginal rate of substitution between expected profit and risk. For the risk-neutral firm, the marginal rate of substitution is zero and for the risk-seeker firm it is negative.[6]

[3]Mathematically we have: (i) $U'(\pi) > 0$ and (ii) $U''(\pi) < 0$ for the risk-averse firm, $U''(\pi) = 0$ for the risk-neutral firm, and $U''(\pi) > 0$ for the risk-seeking firm where $U(\pi)$ is the firm's utility-of-profit function.

[4]The absolute risk-aversion function is written $r_A = -U''(\pi)/U'(\pi)$. Mathematically we have $r_A' > 0$ for the firm with decreasing absolute risk aversion, $r_A' = 0$ for the firm with constant absolute risk aversion, and $r_A' < 0$ for the firm with increasing absolute risk aversion.

[5]This assumption implies that profits are normally distributed. We exclude quadratic utility curves since they display increasing absolute risk aversion.

[6]Mathematically we have $EU(\pi) = f(E, \sigma)$, where E is the expectation operator and σ the standard deviation of profit. The marginal rate of substitution between expected profit (E) and risk (σ) is $dE/d\sigma = -(\partial EU/\partial\sigma)/(\partial EU/\partial E)$. The partial $\partial EU/\partial\sigma$ is negative for risk-averse firms, zero for risk-neutral firms, and positive for risk-seeking firms. The partial $\partial EU/\partial E$ is positive for the three cases. It follows that the marginal rate of substitution is positive for risk-averse firms, zero for risk-neutral firms, and negative for risk-seeking firms.

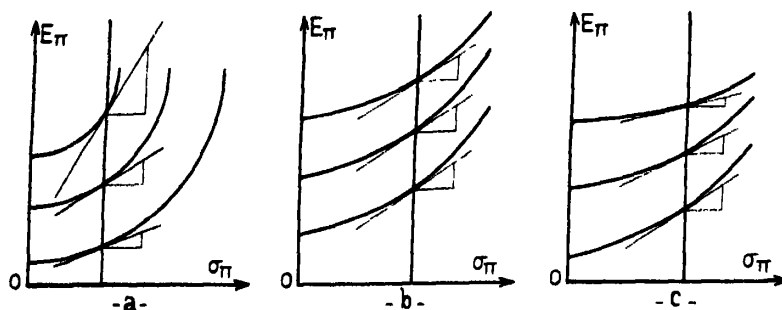



FIGURE 1

C. *The Type of Risk Aversion and the Shape of the Indifference Curves*

The type of risk aversion can be measured by the variation in the firm's marginal rate of substitution when risk is held constant. If, starting from a given level of risk, we move up to higher indifference curves, the same additional unit of risk requires increasing compensating expected profit, then the firm displays increasing absolute risk aversion: its marginal rate of substitution increases with profit for a given level of risk as illustrated in Figure 1a. If the marginal rate of substitution is constant, the firm displays constant absolute risk aversion as shown in Figure 1b. If the marginal rate of substitution decreases with increasing expected utility of profit, the firm displays decreasing absolute risk aversion as in Figure 1c.

III. The Model

A. *The Model under Perfect Competition*

Assume that

$$p = \mu + e \tag{1}$$

where the prices (p) are stochastic and expressed as the sum of the expected price (μ) and a random element (e) with *constant* variance and zero expected value. The *ex ante* prices fluctuate around their expected value and the *ex post* price may differ from μ. Equation (1) implies that

$$E(p) = \mu \tag{2}$$

$$\sigma(p) = \sigma(e) = \text{constant} \tag{3}$$

where E is the expectation operator, $\sigma(p)$ the standard deviation of prices, and $\sigma(e)$ the standard deviation of the random element (e). Firms are further assumed to maximize the expected utility of their profit, that is,

$$\text{Max } E[U(\pi)] \tag{4}$$

where U is the firm's utility function and π the level of profit. The firm's objective function (4) is to be maximized given the constraints, that is, the firm's revenues and costs expressed in a profit function.

The Cost Function. The cost function is assumed to be known with certainty and given by

$$C(q) = V_1(q) + F \tag{5}$$

when (q) is the known level of output, C the total costs, V_1 the variable costs such as $V_1(0) = 0$, and F the fixed costs.

The Profit Function. The profit function (π) is given by

$$\pi = TR - C = p \cdot q - V_1(q) - F \tag{6}$$

where TR is the firm's total revenues.

The Profit-Opportunity Locus. Using equation (6) we can obtain the expected profit and the standard deviation of profit from which the profit-opportunity locus is derived. From equation (6) we have

$$E(\pi) = \mu \cdot q - V_1(q) - F \tag{7}$$

$$\sigma(\pi) = q \cdot \sigma(p) \tag{8}$$

where $E(\pi)$ and $\sigma(\pi)$ are the expected profit

and the standard deviation of profit, respectively. Equation (8) states that $\sigma(\pi)$, which is considered as a proxy for the risk faced by the firm operating under uncertainty, is proportional to the level of output (q). The constant factor is the standard deviation of prices. Since $\sigma(p)$ is known, and since the output q is under the firm's control, it follows that the firm can choose the level of risk $\sigma(\pi)$ it is willing to bear simply by varying the level of output. The profit-opportunity locus is obtained from equations (7) and (8). From equation (8) we have

$$q = \sigma(\pi)/\sigma(p) \tag{9}$$

Substituting in equation (7) we get

$$E(\pi) = \{(\mu/\sigma(p))\cdot\sigma(\pi) - F\} - \{V_2(\sigma(\pi))\} \tag{10}$$

Note that V_2 is the variable cost function in terms of $\sigma(\pi)$ rather than q, and the constant $\sigma(p)$ enters in the coefficients of the function V_2.

B. *The Model under Imperfect Competition*

Under imperfect competition, we assume a demand function $p = f(q) + e$ in which the random disturbance is additive. Assuming that the firm is "quantity setting," the profit function becomes

$$\pi = \{f(q) + e\}\cdot q - V_1(q) - F \tag{11}$$

from which we derive

$$E(\pi) = f(q)\cdot q - V_1(q) - F \tag{12}$$

$$\sigma(\pi) = q\cdot\sigma(e) = q\cdot\sigma(p) \tag{13}$$

since $\sigma(p) = \sigma(e)$. It follows that

$$E(\pi) = h(\sigma(\pi)) - V_2(\sigma(\pi)) - F \tag{14}$$

where the revenue function $h(\sigma(\pi))$ satisfies the condition $h(0) = 0$. Equation (14) is the profit-opportunity locus under imperfect competition for the quantity-setting firm.

IV. Applying the Model: A Comparative Static Analysis

To prove the set of conclusions stated in Section IA, we must subject the model to a comparative static analysis. Starting from an initial equilibrium point, a change in one of the parameters, that is, the expected price μ, the risk $\sigma(p)$,[7] the fixed costs F, or the coefficients of the variable cost function V_2, will lead to a new equilibrium point. A comparison of the initial and final equilibria allows us to reach the desired conclusions.

A. *Comparative Output: Certainty versus Uncertainty* (Conclusion i: Figure 2)

Assuming perfect competition and uncertainty, equation (10) holds: expected profit is equal to the difference between a linear function of $\sigma(\pi)$ with slope $(\mu/\sigma(p))$ and intercept $(-F)$, and the transformed variable cost function V_2 which is assumed to display increasing marginal costs. (This assumption is relaxed in Section IVD.) In Figure 2, $E(\pi)$ is the shaded area between the straight line Ff and the curve V_2.

If certainty prevails, then the profit function becomes

$$\pi = (p\cdot q - F) - V_1(q) \tag{15}$$

In order to compare output under certainty to the uncertain output, p the price under certainty is set equal to μ, and equation (15) is rewritten as the difference between $(\mu\cdot q - F)$ and $V_1(q)$. This is graphed on the same diagram as equation (10) where the vertical axis reads profit π, instead of expected profit $E(\pi)$, and the horizontal axis reads (q) instead of $\sigma(\pi)$. We assume $\sigma(p)$ to be higher than one.[8] Both areas are reproduced in Figure 2 to give the exact shape of $E(\pi)$ and π as a function of $\sigma(\pi)$ and q, respectively, on the mean-standard deviation plane. Superimposing the indifference map on the same plane we can obtain the equilibrium point at A under uncertainty for a risk-averse firm. In the case of risk neutrality, equilibrium is found at point N, the maximum of the expected profit

[7]The risk was defined as $\sigma(\pi)$. However, according to equation (8) any change of the constant $\sigma(p)$ will affect risk as $\sigma(\pi)$.

[8]This assumption is made for expository reasons and does not affect the generality of the results.

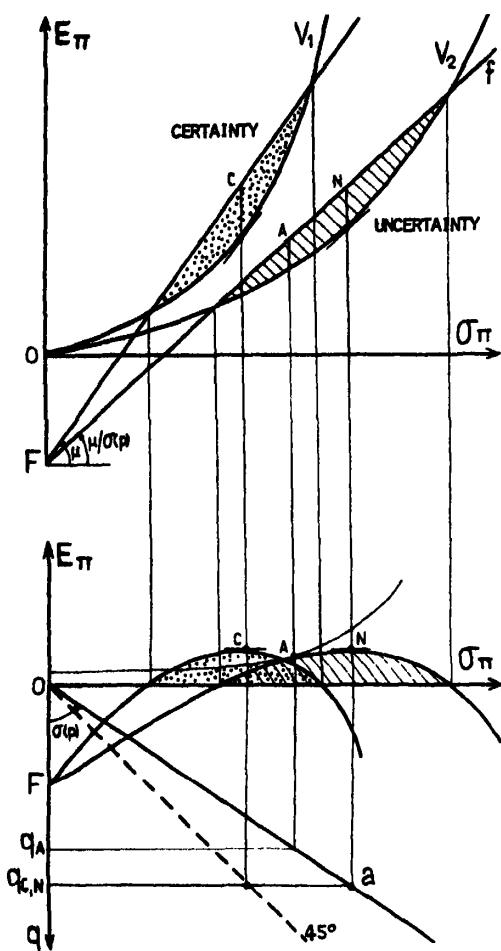



FIGURE 2

curve, since the risk-neutral firm maximizes expected profit. Equilibrium under certainty is found at point C, the maximum of the profit curve.

The Optimum Output. Output can be read along the q axis with output increasing when we move away from the origin downward. The certainty output (q_c) is found using the 45° line, since the $\sigma(\pi)$ axis is the output axis under certainty. Output under certainty and risk neutrality being equal (because risk-neutral firms maximize profit regardless of risk), we can derive the relevant Oa line under uncertainty since we have two points through which the line passes: the origin and point a. Once Oa is

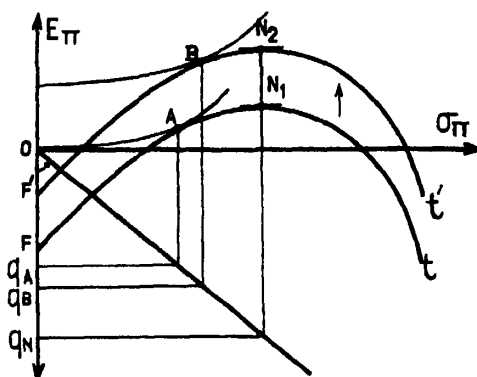



FIGURE 3

drawn we can find output under uncertainty and risk aversion. Since the equilibrium point A is to the left of point N, we obtain a smaller output q_A, which proves Conclusion i under perfect competition and increasing marginal costs. Output under uncertainty and risk aversion is smaller than output under risk neutrality or certainty. The result is independent of the shape of the indifference curve except that risk aversion prevails.

B. *Change in Fixed Costs under Uncertainty* (Conclusion ii: Figure 3)

Assuming perfect competition and increasing marginal costs (later these two assumptions will be relaxed) and referring to Figure 3, we can see that a reduction of

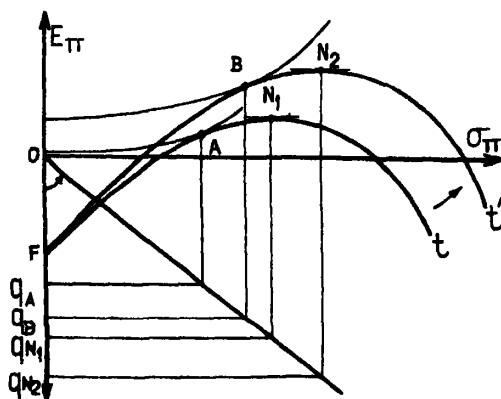



FIGURE 4

fixed costs from OF to OF' shifts up the profit-opportunity locus in a parallel fashion from Ft to $F't'$. Assuming that firms display decreasing absolute risk aversion, the indifference map is of the type described in Figure 1c. The equilibrium point moves from A to B and the corresponding level of output from q_A to q_B with $q_B > q_A$, output varies inversely with fixed costs. For the risk-neutral firm, the equilibrium point moves from N_1 and N_2 with the level of output q_N unaffected. In the case of certainty a similar result is easily derived.

C. *Change in Expected Price and the Supply Curve under Perfect Competition and Uncertainty* (Conclusion iii: Figure 4)

The following assumes that the firm revises its expectations about future prices. An increase in expected price from μ to μ' will rotate the straight line Ff around point F, leaving variable costs constant.[9] As a result, the profit-opportunity locus will shift from Ft to Ft' as indicated in Figure 4. Given decreasing absolute risk aversion, the equilibrium will move from A to B when expected price rises from μ to μ'. The corresponding level of optimum output increases from q_A to q_B. It follows that the decreasing absolute risk-averse competitive firm has an upward-sloping supply curve as in the case of certainty. The conclusion holds true for the risk-neutral firm.

D. *The Cases of Constant and Decreasing Marginal Costs* (Conclusions iv, v: Figures 5, 6)

Referring to Figure 5, observe that an equilibrium exists under constant marginal costs at point A to which correspond the optimum output q_A and the break-even point G.[10] When marginal costs decrease, we move to a new equilibrium point at B

FIGURE 5

with a larger optimum output q_B and a new break-even point D. Under certainty, equilibrium does not exist: firms will maximize their level of output in order to achieve maximum profit (traditional break-even analysis).

FIGURE 6

[9]Referring to Figure 2, we can see that an increase in expected price will rotate the straight line Ff counterclockwise around point F.

[10]In the case of constant marginal costs, the profit-opportunity locus is a straight line.

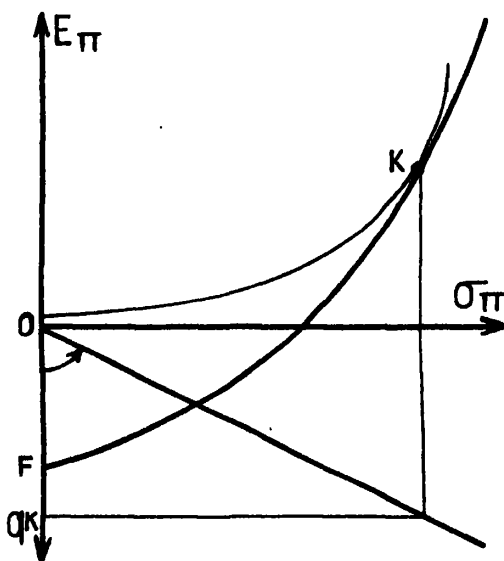
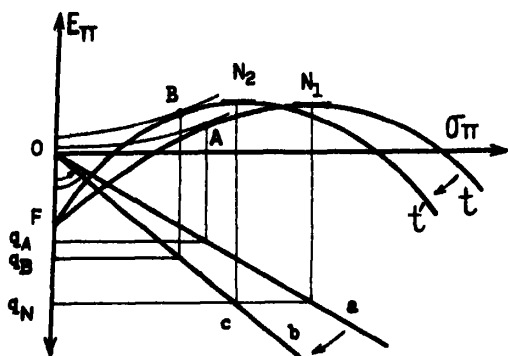



FIGURE 7

Referring to Figure 6, we see that equilibrium may also exist under decreasing marginal costs at point K, to which corresponds the optimum output q_K.[11] Again, under certainty, optimum output does not exist when marginal costs fall. Firms will maximize output.

Referring to Figures 5 and 6 we can see that expected positive profit is required for equilibrium to exist. When the firm expects loss, no equilibrium can be found since the indifference map is not defined in the lower quadrant. This proves Conclusion v.

E. *Change in the Level of Risk*
(Conclusion vi: Figure 7)

The firm can revise its expectations about the distribution of future prices resulting in a revised standard deviation $\sigma(p)$, and therefore a change in the level of risk. Suppose $\sigma(p)$ is revised downward. The profit-opportunity locus will shift from Ft to Ft' in Figure 7 since both the line Ff and the transformed variable costs curve V_2 will rotate counterclockwise around point F as $\sigma(p)$ decreases.[12] The equilibrium point moves from A to B. The corresponding level of output should be derived with caution, since the slope $\sigma(p)$ now varies. Initially the line is Oa and the corresponding output q_A. When $\sigma(p)$ decreases we get a new line Ob which can be easily derived from Oa and the fact that for a risk-neutral firm the level of output q_N remains constant when $\sigma(p)$ changes. This allows us to obtain point c from which the line Ob and output q_B are derived with $q_B > q_A$. It follows that the firm with decreasing absolute risk aversion will increase (decrease) its level of output when risk is revised downward (upward).

F. *Change in Variable Costs*
(Conclusion vii: Figure 8)

Changes in variable costs will leave the line Ff unaffected. Suppose variable costs are reduced, resulting in a rightward shift of the variable cost curve in Figure 2. In this case the profit-opportunity locus will shift upward, rotating around point F as shown in Figure 8. The equilibrium point will move from A to B for the decreasing absolute risk-aversion firm and from N_1 to N_2 for the risk-neutral firm. The level of output will increase from q_A to q_B in the first case and from q_{N_1} to q_{N_2} in the second. As a result of a reduction (rise) in variable costs, the firm with decreasing absolute risk aversion and the risk-neutral firm have increased (decreased) output.

FIGURE 8

[11] In the case of decreasing marginal costs, the profit-opportunity locus is convex to the origin.

[12] Referring to Figure 2, we can see that a decrease in $\sigma(p)$ will rotate both the line Ff and the variable cost curve V_2 counterclockwise around point F resulting in a new profit-opportunity locus Ft' as shown in Figure 7.

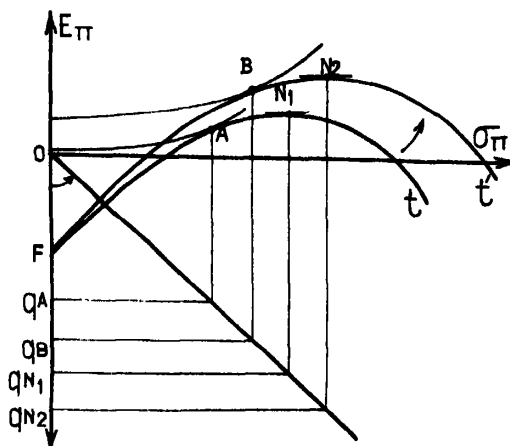
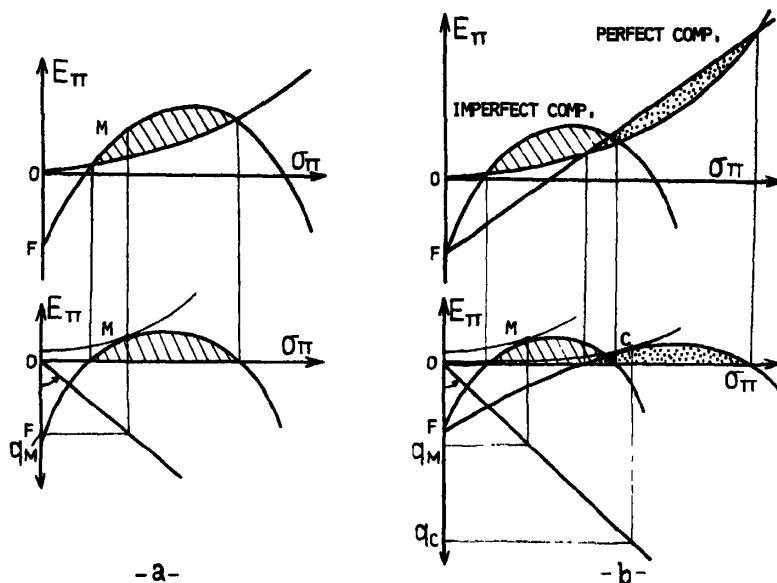



FIGURE 9

G. *Equilibrium under Uncertainty and Imperfect Competition and Comparison with Competitive Equilibrium* (Conclusion viii: Figure 9)

Under imperfect competition, equation (14) holds. Suppose $h(\sigma(\pi))$ is quadratic, that is, the firm faces a linear downward-sloping demand function. Then Figure 9a gives the expected profit area for a firm under noncompetitive market structure. We observe equilibrium at point M and the corresponding optimum level of output q_M. This derivation is similar to that of the competitive firm. It follows that output under uncertainty and imperfect market is smaller than under certainty. This justifies the generality of Conclusion i as well as Conclusions ii, iv, and v since they involve the cost function, which is independent of the market structure.

Comparative Output. Under imperfect competition the slope of the revenue curve $h(\sigma(\pi))$ at point F is larger than the slope of the revenue curve under perfect competition.[13] Referring to Figure 9b, which assumes a similar indifference map for both market structures, we see that the equilibrium point for the imperfect competition case is M under risk aversion, and C under perfect competition and risk aversion. It follows that output under imperfect competition is smaller than output under perfect competition, given risk aversion. A similar conclusion can be drawn for the case of risk

[13]Under imperfect competition expected total revenue is:

(a) $$TR = f\left[\frac{\sigma(\pi)}{\sigma(p)}\right]\cdot\frac{\sigma(\pi)}{\sigma(p)}$$

(b) $$\frac{dTR}{d\sigma(\pi)} = \frac{1}{\sigma(p)}\cdot f'\left[\frac{\sigma(\pi)}{\sigma(p)}\right]\cdot\frac{\sigma(\pi)}{\sigma(p)} + f\left[\frac{\sigma(\pi)}{\sigma(p)}\right]\cdot\frac{1}{\sigma(p)}$$

Therefore $$\lim_{\sigma(\pi)\to 0}\frac{dTR}{d\sigma(\pi)} = \frac{f(o)}{\sigma(p)}$$

where $f(o)$ is the price when output is zero. This limit price is certainly larger than the expected mean price, i.e., $f(o) > \mu$. Under perfect competition expected total revenue is:

neutrality. The level of expected profit corresponding to a given level of output will be higher under imperfect competition than under perfect competition.

(c) $TR = \frac{\mu}{\sigma(p)} \cdot \sigma(\pi)$ and $\frac{dTR}{d\sigma(\pi)} = \frac{\mu}{\sigma(p)}$

Therefore $\lim_{\sigma(\pi) \to 0} \frac{dTR}{d\sigma(\pi)} = \frac{\mu}{\sigma(p)}$

Since $f(o) > \mu$ it follows that the slope of the revenue curve at point F is larger than the slope of the revenue curve under perfect competition.

REFERENCES

H. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 277–91.

Harry Markowitz, "Portfolio Selection," *J. Finance*, Mar. 1952, *7*, 77–91.

———, *Portfolio Selection: Efficient Diversification of Investments*, New York 1959.

A. Sandmo, "On the Theory of the Competitive Firm Under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.

X-Inefficiency Xists—Reply to an Xorcist

By HARVEY LEIBENSTEIN*

Under the title "The Xistence of X-Efficiency," George J. Stigler (1976) wrote a critique of X-efficiency theory, indicated his distaste for the concept, and urged economists to abandon the idea. I will argue that this exercise in exorcism is just that. It achieves some of its results by unusual redefinitions of ordinary concepts. In the end it makes nonscientific appeals. Hence, this plea for exorcism should be ignored. At the same time, I am grateful to Stigler. As a byproduct of his attack he has raised some points that others have raised orally. This provides an opportunity to clarify some issues.

Stigler makes two essential points: 1) X-inefficiency is an illusion, it is really a matter of ignorance and mistakes; and 2) firms minimize costs, despite the fact that they choose different production techniques under similar circumstances. My contention is that X-inefficiency exists. It results from incomplete contracts, effort discretion, and nonmaximizing behavior, rather than lack of information or errors. These basic aspects are developed in Section I below, while Section II considers the problem of choice of technique. Stigler's paper also contains other points with which I disagree. Some of these are handled in Section III.

I. The General Framework for X-Efficiency Theory

It is to be noted that under Stigler's approach, as well as that of traditional theory, X-inefficiency as a phenomenon is simply assumed away. Since neoclassical theory assumes that costs are minimized, X-inefficiency cannot exist. But of course this does not imply that X-inefficiency does not exist in the real world. Hence, to capture the real world phenomenon we must consider a larger theoretical framework than the traditional one. In several papers (1969, 1973, 1975) probably available to Stigler, and in a recent book (1976), I developed such a framework. I shall refer to this larger framework as general X-efficiency theory. Table 1 contrasts the neoclassical model and general X-efficiency theory. Under the latter the neoclassical model can be included as a special case.

With the aid of the concepts developed below I shall try to show that there is nothing in the operation of an economy that is inconsistent with the existence of X-inefficiency, nor does competition necessarily lead to its elimination. Space constraints permit me to do little more than suggest the nature of the basic arguments.

Rational behavior involves complete concern for both the constraints and the opportunities within an economic context. I refer to both of these concerns, that is, constraints and opportunities, as constraint concern. Complete constraint concern is the same as maximization. Selective rationality usually involves less than complete constraint concern. Also, there is a tradeoff between less constraint concern and more internalized *pressure* that an individual feels as a consequence of less concern. Thus an individual's personality will determine the combination of degree of constraint concern and pressure he or she would like to choose—one that he feels most comfortable with. In general the individual strikes a compromise between the way he would *like* to behave (very low constraint concern) and the way he feels he *ought* to behave, which depends on internalized standards for performance and *external* pressures. This implies that individuals do not necessarily or usually pursue gains to be obtained from an opportunity to a maximum degree or marshal information to an optimal degree; also, *maximizing behavior is a special case in this system.* The specific compromise that an individual makes between the competing demands of his id (i.e., unconstrained desires) and his superego (i.e., standard of perfor-

*Harvard University.

TABLE 1

Postulates and Basic Variables	Conventional Micro Theory	General X-Efficiency Theory
1. Psychology	1. Maximization or minimization	1. Selective rationality
2. Firm activity contracts	2. Given	2. Incomplete
3. Units	3. Households and firms	3. Individuals
4. Effort	4. Assumed given	4. Discretionary variable
5. Interpersonal interactions	5. None	5. Some
6. Inert areas	6. None	6. Important variable
7. Agent-principal	7. Identity of interests	7. Differential interests
8. Market structure	8. Given	8. Depends on effort
9. Motivation	9. Implicitly constant	9. Variable

mance), on the average, may be viewed as an index of his personality.[1] However, the pressures determined by a particular context may induce an individual to pursue gains or show greater constraint concern than he would normally find comfortable. Thus personality and context select, so to speak, the *degree* of rationality that will control an individual's decision-making (and performing) behavior.

Our basic decision-making units are individuals even in the case of multiperson households and firms. Further, we assume that firm members have some degree of discretion with respect to the degree of effort they put forth in their work.

Effort is a complex variable which includes such discretionary aspects as the choice of: activities (A), the pace (P) at which activities are carried out, and the quality (Q) and the time pattern (T) with which they are carried out. Within some constraints people choose $APQT$ bundles or effort points. To interpret their jobs, individuals will usually choose a *set* of related effort points in order to be able to meet some differential demands on effort. Also, for each person there is an effort position that contributes to cost minimization, but this is not necessarily the one that will be selected. Since in this system the firm does not control the effort levels of the individuals, it cannot necessarily minimize costs. Note that the deviation between the optimal levels of effort from the firm's viewpoint and the actual level that individuals are motivated to put forth determines the degree of X-inefficiency in the system.

Since motivation is extremely important in determining effort levels, we have to take into account *interpersonal interactions* and, especially, peer group interactions, which determine the system of approval and disapproval. In turn this influences the effort level.

A vital element in our system of analysis is the concept of inert areas—somewhat similar to inertia. Individuals are presumed to choose effort positions (a set of related effort points) to interpret their jobs. The basic idea is that once an effort position (a set of effort points or $APQT$ bundles) exists for some time period, an individual may not shift to a new position, even though there may be a gain achieved thereby because the "inertial cost" of moving from one effort position to another is larger than the perceived gain.[2] We must keep in mind that making (or contributing to) the price, quantity, and quality *decisions* in a firm also requires effort.

[1]The psychological terms id and superego are used in a very general sense and are not to be identified with any psychological system of thought. Substitute words can be found.

[2]Some may argue that maximization is really involved, if we consider the utility of the effort position minus the inertial cost. This is an illegitimate calculation. There is a distinction between person A who maximizes (zero inertial cost) and B, who does not move to a maximum point (positive inertial cost). Clearly A *behaves* as a maximizer and achieves a maximum position which B does not. Inertial cost *explains* why B does not move to a maximum. If A achieves a maximum and B does not, both cannot be called maximizers.

The postulates of incomplete contracts, effort discretion, inert areas, interpersonal influences, and different principal-agent objectives, as outlined, enable us to see why monopolistic firms are likely to have higher costs than the average competitive firms. Start with a firm in isolation; i.e., a firm that is not embedded in a competitive environment. Initially we ignore those motivating elements which would arise out of a competitive environment.

The basic argument follows: Because of incomplete contracts there will exist some degree of effort discretion for firm members. Thus firm members are required to interpret their jobs, which involves the choice of an effort position. Also, different interests between principals and agents, and the existence of effort discretion, imply the possibility of the choice of nonoptimal effort positions; that is, nonoptimal from the viewpoint of cost minimization. Furthermore, the existence of inert areas is consistent with the existence of nonoptimal effort positions which persist over time. Since no one in the firm is presumed to maximize profits, no one is necessarily motivated to try to get the most output from the purchased inputs, and hence costs are not minimized. In other words, under this scheme we would expect X-inefficiency to exist.

Will some degree of competitive pressure result in the reduction or elimination of X-inefficiency? It will lead to the reduction but not necessarily to the elimination of X-inefficiency. Under competition some pressure is put on some members of the various firms whose utility is affected by differential costs and relative profits so that some tightening up (i.e., adjustments) of some effort positions results. This may occur for those who are not in inert areas. For others the pressure may alter their utility functions, which in turn will force some individuals out of their inert areas and induce them to choose more appropriate effort positions in terms of X-efficiency. Additional pressure on existing firms may result from new firms entering or from the expansion of output of existing firms. As the pressures persist, further tightening up of effort positions takes place which results in reduction of costs. But this process may stop short of the point at which X-inefficiency disappears.

In the absence of pressures to contain costs there is a general cost rising tendency. Recall that firm members choose their effort position in the light of some concern for the constraint bounds that exist at the time. However, some of these constraint bounds may be too confining for some employees, and some effort choices go beyond the constraints, which in turn results in pressure against the constraint bounds. Unless managers feel pressures which induce them to struggle against some of the effort choices, especially those that are beyond the "normal" constraint bounds, the result is likely to be rising costs. Thus we may visualize the firm as an arena in which there is a struggle between the cost rising tendencies due to loosening constraint bounds, and the cost containing activities of the management.

A competitive environment may not eliminate X-inefficiency for one of two reasons: There may be a lack of supply of the right kind of entrepreneurs. Assume that there is nothing formal that inhibits entry. Recall that the firm cannot control all costs. An element that determines whether or not there are new entrants is the cost expectations of potential entrepreneurs. Suppose that the X-inefficiency that exists is 20 percent above minimum cost. Now the supply of entrepreneurship may be such that there are no entrepreneurs who believe that they could enter and produce at a lower cost than those already in the industry.

Also, the existing firms may substitute entrepreneurial activities which reduce price competition for minimizing X-inefficiency activities. Let us refer to these as "market-sheltering" activities. They would include cartel creation, monopoly creation, price agreements, product differentiation, arrangements which inhibit entry, and so on. Even if top management were interested in maximizing profits they may nevertheless feel that it pays to substitute some market-sheltering activities for the cost containing struggle to decrease X-inefficiency. In neoclassical theory we do not consider activ-

ities which reduce competition as a substitute for cost-minimizing activities, since cost minimization is assumed to take place, but in the real world this is obviously a possibility. From a theoretical viewpoint there is no reason to exclude this type of substitution or to construct the theory in such a way that we assume the market structure is given.

Another possibility is that an equilibrium may be struck in which new firms enter, but existing firms simultaneously engage in market-sheltering activities which counterbalance the effects of new entrants so that on an overall basis some X-inefficiency persists. Of course, in this case we assume that the new entrants have no preference for cost-reducing activities over market-sheltering activities.

In the light of the previous discussion we have to reconsider what are minimum costs, and what constitutes a competitive environment. The minimum costs idea is straightforward. It is the cost level that would result if firm members attempted to interpret their jobs in such a way that they made effort choices which involved cooperation with peers, superiors, and subordinates, in such a way as to maximize their contribution to output. Obviously, given a system of effort discretion, there is no reason why this should necessarily occur.

By a competitive environment I have in mind a situation in which there are a fairly large number of firms arising as a consequence of no inhibitions to entry. Nevertheless, such an environment need not lead to a situation where there are entrepreneurs always ready and capable of entering the industry in such a way as to achieve minimum costs. This would be a special and unnatural entrepreneurial supply assumption.

Our system accepts the possibility of a state of affairs approximating the neoclassical competitive equilibrium—but only as a special case. In other words, the supply of entrepreneurs willing to operate at minimal cost and capable of doing so may be large enough to counteract the market-sheltering activities. But this is viewed only as a special case which may be approximated in some sectors of the economy. It seems almost obvious that given the elements entering into X-efficiency theory there is absolutely no compelling force for this to be the general case. If anything, actual experience and the arguments presented suggest this would not be so. Effort discretion would reduce this possibility from the viewpoint of the internal organization of the firm. Furthermore, the effort by firm executives to find shelters from the need to struggle against rising costs is an additional element consistent with the persistence of X-inefficiency.

Sometimes evolutionary arguments are put forth to suggest the likelihood that an economy would approximate a zero X-inefficiency equilibrium.[3] The essence of such arguments is that those with minimum costs would survive in the industry, while those with above minimum costs would be forced out. At the same time new firms would enter. Among the new firms, those who would achieve minimum costs would again turn out to be the survivors. The trouble with this argument is its assumption of cost minimization and its persistence. In addition, it assumes that any firm reaching minimum cost will continue to stay there indefinitely. These are very special assumptions. Once we adopt the view that 1) there is a persistent tendency for costs to rise; 2) that in general the cost rising tendency has to be struggled against by management; and 3) that market-sheltering activities are a substitute for cost-reducing activities, the evolutionary argument no longer holds. A balance may be struck between the rising cost tendencies of firms and the struggles against rising costs so that the average costs are above the minimum. The environment is not given. It is created by the firm members who are in the industry. If they choose price-setting and product-differentiating activities which shelter firms from competition, achieving minimum cost ceases to be a survival condition.

[3]See Sidney Winter's reply, and the article by Edith Penrose, to arguments of this sort put forward by Armen Alchian and Milton Friedman.

II. Choice of Technique and Motivation

Stigler's view on choice of technique may be gleaned from the following quotation:

> The near-universal tradition in modern economic theory is to postulate a maximum possible output from given quantities of productive inputs ... and to assert that each firm operates on this production frontier.... The merit of this conventional tradition is also its demerit: it eliminates the problem of the choice of technology.... one may lament ... the failure of Robbins and Leibenstein, and all of us in between to recognize the problem of determining which technologies will be used by each firm.... The choice is fundamentally a matter of investment in knowledge.... Leibenstein deserves credit for revising this Marshallian question, but his attention to X-inefficiency as the explanation is an act of concealment: it simply postulates the differences in technology among firms which should be explained.
>
> [1976, pp. 214 15]

Stigler's assertion to the contrary, I believe that the ideas summarized in the previous section do in fact explain the choice of technique, and that Stigler's denial of the significance of motivation, and the postulate that every firm is on its production frontier, really conceal what actually goes on in the typical multiperson firm.

One of the main notions of X-efficiency analysis is that the technique of production cannot be determined by the choices of some major executive. When the firm hires specific capital and labor inputs it does not determine the actual technique. It is not a matter of mistakes, it is a matter of the extent to which some people can choose what other people do. We argue that effort discretion is a fact of life. Hence, the actual technique must depend on individual choice and on the motivating factors existing both within the firm and in the firm's environment. These in turn strongly influence detailed choices. Thus while the choice of technique from this viewpoint ceases to be mysterious, it is no longer consistent with the assumptions of neoclassical theory.

It is to be noted that Stigler's view of the choice of techniques differs from the standard textbook view in which competing firms have full knowledge, or equal access to knowledge, so that they minimize costs, choose the same technique, and achieve the same minimal cost level (see C.E. Ferguson). Unfortunately, the empirical evidence shows that this is not the case. Hence the theory makes an incorrect prediction. It seems to me that this is a straightforward view of the theory and its results. We should not be inclined to rationalize the deficiencies of the theory, but to face up to some of its limitations.

Stigler argues that all firms operate on the production frontier. However, different firms have different frontiers, possibly because of differences in knowledge, or entrepreneurial capacity—we do not really know. But if every firm has its own frontier, some are further out than others. It is a small step to think of the outermost frontier (or the frontier of frontiers). Hence, some firms are operating inside the outermost frontier, but this does not represent allocative inefficiency. It represents what I have called X-inefficiency. Thus, the idea of X-inefficiency may be hidden by Stigler's approach, but ultimately it is not really avoided. The alternative approach, one that seems more natural, is to say that some firms do not operate on the production frontier. As a consequence differences in costs between firms exist. In my view these cost differences are best explained by the fact that firm members make discretionary effort choices, and that these are influenced by the motivational environment within the firm and between firms.

Stigler argues that motivation is not an input, but part of the problem of choice of technology. But what is the choice of technology problem in the neoclassical sense if it is not to choose the appropriate combination of inputs? The critical question is whether motivation is or is not an input in neoclassical theory, in the real world, or in both? Now, motivation is not an explicit

input in the neoclassical model because the maximization of something can be viewed as assuming away *variations* in motivation. But this does not imply that motivation is not a significant *variable* in the real world.

Despite Stigler's sarcasm contained in his "Romans discovering America" example, there is little doubt that motivation is a significant variable in the real world. There is a large literature in industrial psychology and on business organizations which supports this view. In addition, all of us are aware of situations in which our own motivations changed for some reason, so that we were able to undertake and carry out certain tasks in a superior fashion or more vigorously than would otherwise have been the case. The empirical significance of motivation is not an issue. The real issue is whether we want to think about it in a straightforward manner for the analysis of economic problems. I would suggest that the straightforward approach is to recognize motivation as an obviously important factor in determining productivity, and that frequently we want economic models to contain motivations directly or indirectly as a variable. There may be sets of problems whose analysis is not helped by including variations in motivation. Here the neoclassical theory may be appropriate. However, there is a large set of other problems, especially those for which effort is a discretionary variable, where it makes sense to include motivation as a variable.

In multiperson firms, where effort is a discretionary variable, no one in the firm really controls all that goes on. The nature of the effort depends on the motivation that individuals bring to the firm from their background and personality. They themselves create the motivational environment and structure, since everyone contributes to the interpersonal system of approval and disapproval, which influences the reaction patterns to an individual's effort position choice.

Stigler argues that "the choice [of technique] is fundamentally a matter of investment in knowledge" (1976, p. 215). No doubt knowledge plays a role, but could it carry the burden of the entire explanation? Do individuals and firms acquire all the knowledge it pays for them to acquire? Stigler provides no evidence that this is the case. I believe there is some evidence to the contrary. Stigler speaks of hypothetical farmers making choices in various ways. A recent paper by Kenneth Shapiro and Jurgen Müller examines the choice of technique by farmers in Kenya. It demonstrates that in reality farmers do not use as much knowledge as is available to them, not only because of differences in costs but also because they do not *wish* to. It is not unreasonable to presume that motivation plays a role here. Furthermore, it would be strange to argue that these real farmers are operating on their production frontier.

III. Other Points in Dispute

Early in his paper Stigler argues "Surely no person ever seeks to maximize the output of any one thing: even the single proprietor, unassisted by hired labor, does not seek to maximize the output of corn: he seeks to maximize utility, and *surely* other products including leisure and *health*, as well as corn, enter into his utility function. When more of one goal is achieved at the cost of less of another goal, the increase in output due to (say) increased effort is not an increase in 'efficiency'; it is a *change* in output" (p. 213, emphasis added). This argument does not really handle the issue. It avoids some basic distinctions, that between the product for sale and *intra*firm outputs, and that between single person and multiperson production units.

Consider the three cases in Table 2. It is assumed that the product under consideration is the one sold on the market by a

TABLE 2—ALTERNATIVE POSSIBILITIES OF INCREASES IN COST AND ON-THE-JOB UTILITY (Shown in Percent)

	Cases		
	1	2	3
Δ Cost	+20	+20	+20
Money value of compensating on-the-job utility	0	+10	+20

multiperson firm. Now suppose that some firm members choose to put forth their effort so that the cost of the product is 20 percent higher than minimum cost. In Stigler's language, output is changed if productivity declines. However, I believe it's more convenient to speak of the changes as an increase or decrease in *on-the-job* utility. In Case 1 we see that there is no compensating increase in on-the-job utility for the lower value of effort. In Case 2 we assume that there is some increase in utility but it is worth less than the decreased value of effort. In Case 3 it is assumed that the increased value of utility of the workers is exactly equal to the decreased value of the product. There is nothing in the argument put forth by Stigler to deny the possibility of Cases 1 and 2. Even if in some sense more "leisure" is produced, there is no reason for the value of leisure to be in any sense equal to or greater than the reduced value of the product. Cases 1 and 2 are clearcut cases of X-inefficiency without any counterpart increase in on-the-job utility.

It is even possible to visualize a more extreme case under which the counterpart result of producing less with given inputs can result in a *lower* on-the-job utility than would otherwise be the case. Suppose that two individuals who do not get along are put in leading positions in a department of a firm so that output is lower than would otherwise be the case. This leads to a lower aggregate on-the-job utility than would be the case if things were better arranged and they were in different departments. There is nothing to prevent any of these examples, or something like them, from occurring in real life. We can define things in such a way that such possibilities are assumed out of existence, but this simply uses theory to put blinders on our powers of observation. The main point is that X-inefficiency can exist without complete counterpart increases in on-the-job utility.

There is some discussion about the optimal enforcement of contracts in Stigler's piece, but it is not clear what sort of contracts are being enforced and who does the enforcing. If the effort aspects of a contract are vague, there is not much that can be done to enforce them. What standards of performance should the firm pursue in its enforcement activities if the contract is vague? Furthermore, in a firm run by agents rather than owners, the extent of attempts at enforcement depends on the motivations of the agent-managers. Further, we are told that "... the avoidance of unpleasant tasks and the enforcement activity designed to curtail this avoidance *can* be carried on to the utility-maximizing degree and generate no inefficiency in producing utility" (p. 213, emphasis added). The language used suggests that the enforcement activity, etc., *can be* carried on to a utility-maximizing degree. Even if this *could* be the case, is it the relevant consideration? The question is, will it be? There is no evidence or logic presented to suggest that this will be the result. What is ignored in all this is the discretionary effort gap within which motivation will determine the actual effort level.

Stigler also argues that payments to inputs are adjusted to productivity, and hence this takes into account X-inefficiency. While this may be true in the neoclassical perfect competition model, it does not take account of deviations from perfect competition, nor does it take account of payments to firm members which help to create market sheltering conditions.

One of the empirical questions at issue is whether monopoly leads to higher costs than competition. Stigler provides no evidence to the contrary. There is some evidence on this matter. Walter Primeaux, Jr. compared the forty-nine cities with more than one electric power company with those that have single electric power companies. After taking economies of scale into account, Primeaux found that

> ... on the average, cost is reduced, at the mean, by 10.75% because of competition. This reflects a quantitative value of the presence of X-efficiency gained through competition; or an estimate of the loss caused by the absence of competition in a regulated environment. [p. 107]

The value of the X-efficiency hypothesis is reflected in two other studies. John Shel-

ton compared owner-operated franchise restaurants with manager-operated units. Despite the very high degree of standardization of menus, accounting systems, etc., and despite the virtual equality of sales volume, the owner-operated units averaged a profit margin of 9.5 percent but the manager-operated units averaged only 1.8 percent. T.Y. Shen studied technological diffusion as an element that influenced cost in 4,000 plants. He found that

> ... it is necessary to recognize the presence of a further systematic influence that also affects the change of input-output combinations of manufacturing plants over time ... We find the observed behavior pattern is better explained by the prevalence of 'X-efficiency' rather than by substitution ... This tentative finding is put to a further test by a perusal of the nature of a factor intensity change. Once more the X-efficiency hypothesis turns out to be more consistent with the data. We conclude that a technological change model based on diffusion requires the estimation and incorporation of X-efficiency. Until this step is taken, the use of the extended diffusion model for explaining growth is of dubious validity. [p. 264]

These represent some studies which show the value of X-efficiency theory over its strict neoclassical counterpart.

Since I have postulated a nonmaximizing framework in which maximization is a special case, some remarks on maximization are in order. The maximization hypothesis to be meaningful must be falsifiable. That is, if the word "maximum" means something, then the possibility of choosing a nonmaximum must also exist. Hence the tautological approach to maximization eliminates meaningfulness. In the third edition of Stigler's justly famous text (1966) he seems to argue in favor of the tautological interpretation of utility maximization. He defends it on the ground that at the very least this approach leads to accurate predictions. However, Stigler himself does not indicate what these predictions are. Consider the case of voting behavior. If people really maximized utility, no one would vote except those who felt that they had a deciding vote, and were concerned about the outcome. The evidence shows that in fact the majority of the population frequently does vote, although on the average a higher proportion will vote when the election is close or strong private interests are involved. This is just one of many examples of this kind. In other words, the maximization postulate can be false in many instances. Nevertheless, my general view is that in some cases the maximization model may be a useful simplification, but we should not attempt to use it everywhere.

In general Stigler appeals to tradition and warns against "the mighty methodological leap into the unknown that a non-maximizing theory requires" (p. 216). Both the appeal and the warning are nonscientific. After over a century of maximizing models, it is time to consider nonmaximizing approaches.

REFERENCES

A. A. Alchian, "Uncertainty, Evolution, and Economic Theory," *J. Polit. Econ.*, June 1950, *58*, 211-21.

Y. Barzel and E. Silverberg, "Is the Act of Voting Rational?," *Publ. Choice*, Fall 1973, *16*, 51-58.

C. E. Ferguson, *Microeconomic Theory*, 3d ed., Homewood 1972.

M. Friedman, "The Methodology of Positive Economics," in *Essays in Positive Economics*, Chicago 1953.

H. Leibenstein, "Aspects of the X-Efficiency Theory of the Firm," *Bell J. Econ.*, Autumn 1975, *6*, 580-606.

———, "Organizational or Frictional Equilibria, X-Efficiency and the Role of Innovations," *Quart. J. Econ.*, Nov. 1969, *83*, 600-23.

———, "Notes on X-Efficiency and Technical Progress," in Eliezer B. Ayal, ed., *Micro Aspects of Development*, New York 1973.

———, *Beyond Economic Man*, Cambridge, Mass. 1976.

E. T. Penrose, "Biological Analogies in the Theory of the Firm," *Amer. Econ. Rev.*,

Dec. 1952, *42*, 804–19.

W. J. Primeaux, Jr., "An Assessment of X-Efficiency Gained through Competition," *Rev. Econ. Statist.*, Feb. 1977, *59*, 105–07.

K. H. Shapiro and J. Müller, "Sources of Technical Efficiency: The Roles of Modernization and Information," *Econ. Develop. Cult. Change*, Jan. 1977, *26*, 293–310.

J. P. Shelton, "Allocative Efficiency vs. X-Efficiency: Comment," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1252–58.

T. Y. Shen, "Technology Diffusion, Substitution, and X-Efficiency," *Econometrica*, Mar. 1973, *41*, 263–84.

George J. Stigler, *The Theory of Price*, 3d ed., New York 1966.

———, "The Xistence of X-Efficiency," *Amer. Econ. Rev.*, Mar. 1976, *66*, 213–16.

S. G. Winter, "Satisficing, Selection, and the Innovating Remnant," *Quart. J. Econ.*, May 1971, *85*, 237–61.

A Theory of Employee Job Search and Quit Rates

By KENNETH BURDETT*

The purpose of this study is to develop and analyze a model of job search which allows for the possibility of workers looking for a job while employed. Such a model leads to significant generalizations of results already in the job search literature. Results obtained are then used to specify a theory of job quits and quit rates.

Previous studies on job search, with one recent exception[1] (see Donald Parsons, 1973), assume workers never look for a job while employed. James Tobin has noted that this restriction can only be justified if job search is significantly more efficient when unemployed, and this is patently not the case in many actual labor markets.[2] In support of this argument J. Peter Mattila has estimated that 60 percent of those workers who voluntarily change jobs in the United States suffer no interim unemployment. Since this can only occur if some employed workers obtain new jobs before quitting, allowing workers to search while employed appears to be valid empirically. The consequences of such a change are of some importance as job search models are at the center of much modern work on unemployment and inflation.

The basic structure used in job search models is now well known.[3] Unemployed workers select a strategy to maximize their own discounted lifetime income in a market where job offers are envisaged as random draws from a known distribution of wage offers. It has been shown that the best strategy in such a market is for a worker to select a reservation wage before an offer is received. Any offer then made to that worker will be accepted if and only if the wage offered is at least as great as the reservation wage. If an offer is accepted, the worker is assumed to work at the firm until retirement, presumably because the cost of looking for a better job is too great. In this study it is shown that this is only the best strategy if the cost of looking for a job while employed is high relative to the cost when unemployed. If this is not the case another strategy yields a greater expected payoff. This strategy involves a worker selecting two "reservation" wages, X and Y, where $X < Y$. An unemployed worker will then accept any offer if and only if the wage offered is at least as great as X. However, if the wage offered is acceptable but less than Y, the worker will continue to look for another job when employed. Any offer with a wage at least as great as Y implies the worker will accept and not look for a job when employed. An employed worker who is looking for another job will accept any offer received with a wage greater than his current wage. Details of these results are presented in Section I.

Two facts dominate the empirical literature on quit rates: the probability a worker quits a job declines as the worker's age increases, and as job tenure increases.[4] Two explanations of these results have been proposed. First, it has been argued that workers accumulate firm-specific capital as they work at a firm (see Parsons, 1972). This implies the wage of a worker will increase relative to the next best alternative as tenure increases if workers are paid according to their marginal products. The second explanation stresses the idea that workers do not know all the relevant characteristics of a firm when becoming employed (see Dale

*Department of economics, University of Wisconsin-Madison. I would like to acknowledge the help given by Dale Mortensen. I am responsible for any remaining errors.

[1]To generate results on employee job search Parsons assumes that the cost of search function is quadratic. No such restriction will be used in the present study.

[2]Tobin uses the example of the labor market for academic economists as one where the cost of search when employed is not greater than the cost when unemployed.

[3]Steven Lippman and John McCall present an excellent survey of job search studies.

[4]Robert Hall's article is a good example of this literature.

Mortensen, 1975). A worker in this case may decide to quit if the characteristics of the firm learned when working at the firm make the job unacceptable. Using either of these explanations it is possible to derive a negative relationship between quitting and job tenure. The negative relationship between quitting and age follows as a consequence of the positive correlation between age and job tenure.

In this study workers do not accumulate firm-specific capital and know all about a job before starting employment. Workers quit only because a better wage offer is found. Quits of this type may be termed *wage quits*. There are two possible causes of such behavior. First, a worker's wage may decline relative to others. Quits motivated by such a change are termed *dynamic wage quits*. In the present study workers look for another job when employed as part of an optimal search strategy even when relative wages are held constant. Quits of this type may be termed *equilibrium wage quits*. This type of quit can occur as a long-run feature of a market, whereas dynamic wage quits only occur after some shock to the system. With equilibrium wage quits the causal relationship is between quitting and age. The association between quitting and job tenure follows as a consequence of the positive correlation between tenure and age. The job quits theory developed leads to the prediction that the average wage received by workers of a given age increases as the age group considered increases. This is similar to results in the human capital literature. The reasons behind this prediction are different, as workers do not accumulate human capital (by assumption).

I. Formal Model of Job Search

The labor market structure outlined in this section is similar to that utilized in most previous job search studies. Its most noticeable feature is that all firms do not offer the same wage rate in any given period. Rather it is assumed that the wage offers made by all firms in the market can be described by a nondegenerate distribution function $F(w)$. Associated with this function is a density function $f(w)$.

Suppose a worker's working life can be divided into N periods. In any period a worker can pay a fixed amount (the search cost) and receive one job offer. As workers are assumed not to know which firms are offering which wage, an offer can be envisaged as a random draw from the known distribution of wage offers. This implies $(1 - F(w''))$ denotes the probability a worker who attempts to obtain a job in a period receives an offer with a wage at least as great as w''. A worker who accepts an offer works at the wage rate offered per period until he retires or quits. If an offer is rejected, the worker is assumed capable of returning to accept it in a later period. An unemployed worker who attempts to obtain a job in a period is eligible for unemployment insurance payment u in that period. Unemployed workers who do not look for a job are ineligible for such a payment.

Suppose workers attempt to maximize their own expected discounted lifetime income net of search costs. In previous job search studies it has been assumed a worker selects one of two options in any period: *option* 1, look for a job (search) but not work; or *option* 2, work but not search. In this study a worker will be allowed to choose a further option: *option* 3, work and search. The cost of looking for a job while employed may be different from the cost while unemployed. For example, the cost of search when employed may include loss of earnings while searching. To allow for such possibilities, let c_1 and c_2 denote the cost of search when unemployed and employed, respectively.

Although only three options will be considered in the present study, there is another option open to workers in many labor markets. With this option a worker selects to neither work nor search. This has been termed the discouraged worker option in the literature. To simplify the exposition the discouraged worker option will be ruled out by assumption. A simple way of achieving this goal is to assume $u - c >$ 0. It is straightforward to check a worker

will always prefer option 1 to the discouraged worker option if this restriction holds. Ruling out this option does not imply it is not relevant in many market situations but reflects my desire to concentrate on other issues.

A worker about to begin period t of working life will be said to be of (working) age t. Suppose a worker of age t has received a maximum wage offer w'. Let $\mu_{1t}(w', u, c_1)$ denote the maximum expected discounted lifetime income to this worker given that option 1 is selected in the next period and then an unrestricted choice of options is allowed in all future periods. Similarly, let $\mu_{2t}(w')$ indicate the maximum expected payoff if option 2 is chosen in period t and $\mu_{3t}(w', c_2)$ if option 3 is selected when an unrestricted choice is allowed in all future periods. The worker will choose the option in a period that yields the greatest expected discounted lifetime income net of search costs. Let

$$(1)\quad \psi(w', t) = \max\{\mu_{1t}(w', u, c_1), \mu_{2t}(w'), \mu_{3t}(w', c_2)\}$$

denote the maximum expected payoff to a worker of age t who has received a maximum wage offer w' to date. The expected payoffs to choosing each of these three options in period t of working life when w' is the best wage offer received can be written as

$$(2)\quad \mu_{1t}(w', u, c_1) = u - c_1 + \beta(1 - F(w'))E\{\psi(w, t+1) \mid w \geq w'\} + \beta F(w')\psi(w', t+1)$$

$$(3)\quad \mu_{2t}(w') = w' + \beta\psi(w', t+1)$$

$$(4)\quad \mu_{3t}(w', c_2) = w' - c_2 + \beta(1 - F(w'))E\{\psi(w, t+1) \mid w \geq w'\} + \beta F(w')\psi(w', t+1)$$

for any $t < N$, where $\beta = 1/(1+r)$ is the discount rate and r the discount factor. What option will a worker choose in the last period before he retires? Without any real loss of generality assume $u - c_1 < w'$ for any possible wage w' in the market. This ensures that a worker will always select option 2 (work and not search) in period N in the market, that is,

$$(5)\quad \psi(w', N) = \mu_{2N}(w') = w' \quad \text{for any possible wage offer } w'$$

This claim is simple to check if it is noted $\psi(w', N+1) = 0$ for any w'.

The expected payoff to selecting any option in a period depends on the maximum wage offer received at the start of the period. Taking the partials of (2), (3), and (4) with respect to w' and using (5) implies

$$(6)\quad \frac{\partial \mu_{2t}(w')}{\partial w'} > \frac{\partial \mu_{3t}(w', c_2)}{\partial w'} > \frac{\partial \mu_{1t}(w', u, c_1)}{\partial w'} > 0$$

when evaluated at any given w' and for any $t < N$. Hence an increase in w' to $w' + \delta$ ($\delta > 0$) will increase the expected payoff to selecting option 2 in the next period more than the payoff to option 3, which in turn increases more than option 1.

When will option 1 be preferred to option 3 in any given period? From (2) and (4) it follows that if the maximum wage offer received to date equals z, where

$$(7)\quad z = u - c_1 + c_2$$

then the expected payoffs to options 1 or 3 in the next period are the same, i.e., $\mu_{1t}(z, u, c_1) = \mu_{3t}(z, c_2)$ for any t. Further, (6) implies

$$(8)\quad \mu_{1t}(w', u, c_1) \gtreqless \mu_{3t}(w', c_2) \qquad \text{as } w' \lesseqgtr z$$

Thus a worker will select option 3 in preference to option 1 in any period if and only if the maximum wage offer received to date is at least z. The situation is not so simple when other pairs of options are compared. Consider options 1 and 2. For any fixed $t < N$ let x_t denote the maximum wage offer received to date that equates the payoffs to selecting either of these options in period t of working life. Equating (2) with (3) yields

$$(9)\quad x_t = u - c_1 + \beta \int_{x_t}^{\infty} \{\psi(w, t+1) - \psi(x_t, t+1)\} f(w)\, dw$$

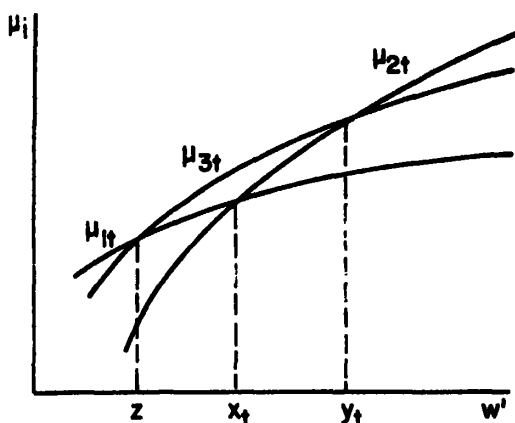



FIGURE 1a

FIGURE 1b

Similarly, let y_t indicate the maximum wage offer received to date that equates the payoffs to choosing options 2 or 3 in any period t of working life. This implies

$$(10)\quad c_2 = \beta \int_{y_t}^{\infty} \{\psi(w, t+1) - \psi(y_t, t+1)\} f(w)\,dw$$

Using (6) it can be seen both x_t and y_t are unique for any fixed $t < N$ and

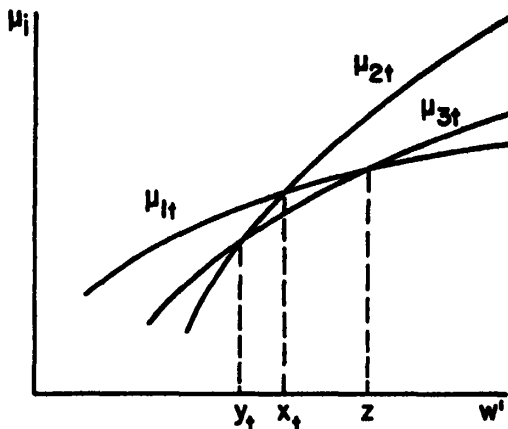
$$(11a)\quad \mu_{1t}(w', u, c_1) \gtrless \mu_{2t}(w') \quad \text{as } w' \lessgtr x_t$$

$$(11b)\quad \mu_{3t}(w', c_2) \gtrless \mu_{2t}(w') \quad \text{as } w' \lessgtr y_t$$

An important consequence of the above analysis is presented in the following claim.

PROPOSITION 1: *$z \gtrless x_t$ if and only if $y_t \lessgtr x_t$.*
The proof is shown in the Appendix.

Suppose $z \leq x_t$. This situation is depicted in Figure 1a. As a worker will select the option in period t that maximizes expected payoff, it can be seen by inspection of Figure 1a that the following strategy is optimal in period t of working life.

Strategy A:
(a) Select option 1 (search not work) in period t only if $w' < z$
(b) Select option 2 (work not search) in period t only if $w' \geq y_t$
(c) Select option 3 (work and search) in period t only if $z \leq w' < y_t$

where z and y_t are defined in (7) and (10) and w' is the maximum wage offer received at the start of period t of the worker's working life.

Suppose $z > x_t$. This situation is depicted in Figure 1b. In this case the optimal strategy for a worker of working age t is as follows.

Strategy B:
(a) Select option 1 (search not work) in period t only if $w' < x_t$
(b) Select option 2 (work not search) in period t only if $w' \geq x_t$

where x_t is defined in (9) and w' is the maximum wage offer received.

Note that x_t is irrelevant for decision purposes if Strategy A dominates, whereas z and y_t are irrelevant if Strategy B is preferred. The next claim is a direct consequence of the above analysis and therefore no proof is presented.

PROPOSITION 2: *For a worker about to begin period t of working life, Strategy A maximizes expected discounted lifetime income only if $z \leq x_t (z \leq y_t)$; otherwise Strategy B is preferred.*

In previous job search models only options 1 or 2 were allowed to be selected by a worker in any period. This restriction implies Strategy B must be chosen by a worker in each period. The more general model considered in this study has reached a more general conclusion. Specifically, it has been

shown that Strategy B is preferred by a worker of age t only if $z > x_t$.

So far the optimal strategy of a worker in any given period has been analyzed. In the remaining part of this section it is shown how the switchpoint wages x_t and y_t change as a worker becomes older. Of course, the wage z is independent of the age of a worker. Given (1)–(6) it is possible to solve z, x_t, and y_t for all $t \leq N$. Proposition 3 summarizes the results that can be obtained from such a task.

PROPOSITION 3:
(a) $y_{t-1} > y_t$ *if Strategy A is preferred in period t of working life*
(b) $x_{t-1} > x_t$ *if Strategy B is preferred in period t of working life*
(c) *If Strategy B is preferred by a worker in period t^* then strategy B will be preferred by that worker in any period $t > t^*$*
The proof is shown in the Appendix.

Claims (a) and (b) establish that the relevant switchpoint wages x_t or y_t decrease as the worker becomes older.[5] An important implication of this result is that workers will only quit a job to become employed at a firm that has offered a better wage. Workers will not quit a job for unemployment in the environment specified.

Several alternative formulations of the model can be considered without disturbing the basic results. For example, it can be assumed that the cost c_2 is a positive function of the worker's current wage, or that a worker cannot return to a previously rejected offer, without radically altering the results obtained. Similar results are also achieved if workers are assumed to have infinite working lives, although in this case the switchpoint wages y_t and x_t remain constant as t changes. If the distribution of wage offers changes through time, however, significantly different results accrue. For example, in this case a worker may choose to quit a job to become unemployed.

[5]Another consequence of this claim is that the switchpoint wages y_t and x_t can be written as

$$c_2 = P(t+1)\beta \int_{y_t}^{\infty} (w - y_t) f(w)\,dw$$

$$x_t = P(t+1)\beta \int_{x_t}^{\infty} (w - x_t) f(w)\,dw$$

where $P(t+1) = \sum_{i=1}^{N-t} \beta^i$

II. Quit Rates

In this section simple job quit functions are investigated. It is first shown that the results of the previous section imply a functional relationship between quitting and (working) age. This implication is then used to show that the probability of quitting decreases as job tenure increases. It should be noted that only equilibrium wage quits as defined earlier are considered. There are many other reasons a worker may quit a job, but these will be ignored to highlight the major argument.

Within the context of the model developed in this study, a worker will quit a job if and only if a better wage offer is received in that period. This event can only occur if option 3 (work and search) is chosen in that period. Workers who utilize Strategy B, however, never select option 3 and thus never quit. Indeed, as previous job search studies assumed workers do not look for another job when employed, an implication of these studies is that workers never quit. To simplify the exposition in this section it will be assumed that $y_t \geq z (y_t \geq x_t)$ for all $t \leq N$. This assumption and Proposition 3 guarantee all workers use Strategy A in each period. Another consequence of this restriction is that only those in their first period in the labor force will be unemployed. To establish this last claim, note from (5) that $F(y_N) = 0$. This result and the restriction $y_t \geq z$ imply $F(z) = 0$, which leads directly to the stated claim from the definition of Strategy A. Although it is somewhat unrealistic to make assumptions that guarantee all workers who have been in the market more than one period are employed, it will not affect the results on job quits and much simplifies the analysis.

Assumptions made above ensure that all workers who have been in the market more

than one period select either option 2 or 3 in any period. From the definition of Strategy A it follows that a worker of age t employed at wage w'' will choose option 3 in the next period only if $w'' < y_t$. Note that as $y_{t-1} > y_t$ (from Proposition 3), a worker who once selects option 2 will choose this option until retirement. A worker in his first period in the market is forced to select option 1 (search not work) as a job offer will not have been received. Define $t(w'')$ as the maximum t such that $y_t \geq w''$ for any fixed w''. This implies that of all workers who work at wage w'' only those with no greater than $t(w'')$ will work and search in the next period. If a worker of age t employed at w'' selects option 3 in a period, there is a probability $(1 - F(w''))$ a better offer is received and hence the worker quits. Let $q(w'', t)$ denote the probability a worker of age t employed at wage w'' quits during the next period. It follows that

$$(12) \quad q(w'', t) = \begin{cases} 1 - F(w'') & \text{if } t \leq t(w'') \\ 0 & \text{if } t > t(w'') \end{cases}$$

A worker employed at a wage less than another worker of the same age is more likely to quit (or at least not less likely) for two reasons. First, if both select option 3 in a period, the lower paid worker is more likely to obtain a greater wage offer than current wage, i.e., $1 - F(w'') > 1 - F(\hat{w}'')$ if $w'' < \hat{w}''$. Second, the lower paid worker will choose option 3 at least as long as the other worker, i.e., $t(w'') \geq t(\hat{w}'')$ if $w'' < \hat{w}''$. These predicted relationships between age, current wage, and the probability of quitting appear not to be refuted in the empirical studies on the subject (see Hall).

Above it was argued that the wage rate and age of a worker determine the probability of quitting. The worker's job tenure did not influence the quitting decision. Nevertheless, it will be shown the above results imply a relationship between quitting and job tenure. This goal is achieved in three stages. First, the expected number of each age group who search a firm in a period is specified. Second, the expected number of each age group who accept a firm's offer is calculated. Finally, it is shown how many quit or retire from a firm in a period after s periods of employment at the firm. The last result leads directly to the claimed negative relationship between quitting and job tenure.

Suppose K new workers enter the market each period. Due to the assumptions made all these workers will accept the first offer they receive. Some of the workers will obtain an offer with a wage at least as great as y_2. These workers will choose option 2 in each of their future periods in the market. Those that received a wage offer less than y_2 in their first period in the market will select option 3 in the next period. Hence the expected number of workers in period 2 of their working lives who choose option 3 is denoted by $KF(y_2)$. Let k_t indicate the expected number of age t workers who select option 3 in a given period. It follows that

$$k_t = K[F(y_t)]^{t-1} \quad \text{if } 1 \leq t \leq N-1 \quad \text{and} \quad k_N = 0$$

Assume γ percent of each working age group who are looking for a job in a period visit a particular firm. This is a harmless restriction in the model considered, as it has previously been assumed that workers randomly choose the firm to search in a period from the set of all firms. This implies γk_t denotes the expected number of workers of working age t who search a particular firm in a period. Note that the expected number who search a firm in a period is independent of the wage offered by that firm. This is not the case when considering how many workers accept a firm's offer.

Consider a firm offering wage rate $\tilde{w}$ in each period. Of those workers searching this firm in a period only those who have not previously received a better offer will accept. This implies that the expected number of age t who accept the offer of this firm in a given period N_t can be written as $N_t = \gamma k_t Pr$ (best offer in $t-1$ offers $< \tilde{w} \mid$ best offer in $t-1$ offers $< y_t$) $t = 1, 2, \ldots, N-1$. Using a standard result from prob-

ability theory yields

$$(13)\quad N_t = \begin{cases} \gamma k_t[F(\tilde{w})]^{t-1} & \text{if } t < t(\tilde{w}) \\ \gamma k_t[F(y_t)]^{t-1} & \text{if } t \geq t(\tilde{w}) \end{cases}$$

Let λ_t indicate the proportion of those who begin to work for a firm in a given period who are working age t. Note that workers who accept an offer in period $t-1$ of working life do not start to work at the firm until they are working age t. The next proposition uses this fact and (13) to specify how λ_t changes with t. As the proof follows directly from the analysis above none will be presented.

PROPOSITION 4:

$$\lambda_t = \lambda[F(e_t)]^{t-2} \qquad t = 2, 3, \ldots, N$$

where $e_t = \min\{y_{t-1}, \tilde{w}\}$

and λ *is chosen so that* $\sum_{t=2}^{N} \lambda_t = 1$

Proposition 4 specifies the expected working-age distribution among the group who accept a firm's offer in a given period. This result will now be used to calculate the expected proportion of the group who start to work for a firm in a given period, who quit or retire from the firm s periods later. First, consider the expected number who quit in each period. From (12) it follows that of those who enter this firm in a given period, a proportion $\lambda_2 + \lambda_3 + \ldots + \lambda_{t(\tilde{w})}$ will select to work and search in their first period at the firm. This implies that $(1 - F(\tilde{w}))\{\lambda_2 + \lambda_3 + \ldots + \lambda_{t(\tilde{w})}\}$ denotes the expected proportion who quit after one period at the firm. Workers of working age $t(\tilde{w})$ when entering this firm who fail to obtain a better offer in the first period select not to search in their second period at the firm as $y_{t(\tilde{w})+1} < \tilde{w}$ by definition. Hence, $F(w)\{\lambda_2 + \lambda_3 + \ldots + \lambda_{t(\tilde{w})-1}\}$ indicates the expected proportion who choose option 3 in their second period after entering the firm. Continuing in a similar way it is possible to specify $Q(\tilde{w}, s)$, the expected proportion of those who enter a firm in a given period who quit that firm s periods later. Proposition 4 and (12) imply

(14)

$$Q(\tilde{w}, s) = (1 - F(\tilde{w}))[F(\tilde{w})]^{s-1} \sum_{i=2}^{t(\tilde{w})-s+1} \lambda_i$$

Manipulating (14) yields[6]

$$(15)\quad Q(\tilde{w}, s) = \lambda[F(\tilde{w})]^{s-1}(1 - [F(\tilde{w})]^{t(\tilde{w})-s}) \quad \text{if } s < t(\tilde{w})$$

and $\quad q(\tilde{w}, s) = 0$ if $s \geq t(\tilde{w})$

After N periods in the market, workers retire. Let $R(\tilde{w}, s)$ indicate the expected proportion of those who enter the firm in a given period who retire while still employed at that firm s periods later. Proposition 4 and (15) imply

$$(16)\quad R(\tilde{w}, s) = \begin{cases} \lambda[F(y_{N-s})]^{N-s-1} & \text{if } s < N - t(\tilde{w}) \\ \lambda[F(\tilde{w})]^{t(\tilde{w})-1} & \text{if } N - 1 \leq s \leq N - t(\tilde{w}) \\ 0 & \text{if } s > N - 1 \end{cases}$$

Finally, let $V(\tilde{w}, s)$ denote the expected proportion of the group who entered the firm in a given period who are still employed at that firm s periods later. As $V(\tilde{w}, 0) = 1$, it follows that

$$(17)\quad V(\tilde{w}, s) = 1 - \sum_{j=1}^{s} \{Q(\tilde{w}, j) + R(\tilde{w}, j)\}$$

After the above tedious but necessary derivation of stocks and flows, the major result of this section can be stated.

PROPOSITION 5: *The quit rate,* $Q(\tilde{w}, s)/V(\tilde{w}, s-1)$, *declines as s increases if* $Q(\tilde{w}, s)$ *is strictly positive.* $Q(\tilde{w}, s)$ *is strictly positive if* $s < t(\tilde{w})$. *Further,* $Q(\tilde{w}, 1) > Q(\tilde{w}', 1)$ *and* $Q(\tilde{w}', s^*) > 0$ *for any* s^* *implies* $Q(\tilde{w}, s^*) > 0$ *if* $\tilde{w} < \tilde{w}'$.

The proof is shown in the Appendix.

[6]The following result is used in this manipulation. If $0 < x < 1$ and M is any positive integer, then

$$\sum_{i=1}^{M} x^i = (x(1 - x^M))/(1 - x)$$

In the remaining part of this study the relationship between the distribution of wage rates received by each working age group is investigated. Let $H_t(w)$ denote this distribution for workers of age t. Hence $H_t(w'')$ indicates the proportion of workers of working age t who are employed at a wage no greater than w''. A worker in period t of working life will receive a wage at least as great as the wage received in period $t - 1$ of working life. Some workers who select option 3 in period $t - 1$ of their working lives will obtain a better offer and hence receive a greater wage in period t. This implies $H_{t-1}(w'') > H_t(w'')$ for any w'' for any $t \leq N$ if $0 < F(w'') < 1$. Using this result and Proposition 4 it can be seen that a randomly selected worker of a given age is more likely to quit in the next period than another randomly selected worker of a greater working age. This claim follows if it is noted that (a) a worker receiving wage w'' of working age t is no more likely to quit in the next period than another receiving the same wage of working age $t + 1$; and (b) more workers of working age $t + 1$ receive a wage at least as great as w'' than workers of age t, for any possible w''.

An implication of the above analysis is that the average wage received by workers of a given age increases as the age group considered increases. Further, it is straightforward to show that this increase in average wage is at a decreasing rate. Formally, we have

$$E\{H_{t-1}(w)\} < E\{H_t(w)\}$$

and

$$|E\{H_{t-1}(w)\} - E\{H_t(w)\}| > |E\{H_t(w)\} - E\{H_{t+1}(w)\}|$$

for any $t < N$. This prediction is similar to that often made from human capital theory, and one that appears to be valid empirically. The reason behind this prediction, however, is quite different. The basic idea behind the human capital explanation is that older workers receive higher wages, on average, because they have accumulated more human capital while working. Hence, if two workers attain the same educational level, the older worker will on average receive a greater wage rate as job experience increases productivity. In the present study it has been assumed workers do not accumulate human capital while working. Older workers in the present study receive higher wage rates, on average, because they have obtained more job offers. And the more job offers a worker receives, the greater the probability a "high" wage rate job will be found.

APPENDIX

PROOF of Proposition 1:

Suppose $z > x_t$. From (8) it follows $\mu_{1t}(x_t, u, c_1) > \mu_{3t}(x_t, c_2)$. Hence $\mu_{1t}(x_t, u, c_1) > \mu_{2t}(x_t)$ from (11a), which implies $x_t > y_t$.

Suppose $x_t > y_t$. From (11b) it follows $\mu_{2t}(x_t) > \mu_{3t}(x_t, c_2)$. Hence $\mu_{1t}(x_t, u, c_1) > \mu_{3t}(x_t, c_2)$ from (11a), which implies $z > x_t$. The rest of the claims of the proposition are established in a similar fashion.

PROOF of Proposition 3:

Suppose $x_t \geq z$ and hence from Proposition 1, $y_t > x_t$. Proposition 2 implies that Strategy A dominates and $\mu_{2t}(y_t) = \mu_{3t}(y_t, c_2) = \psi(y_t, t)$. The equations of (11) imply that if $\mu_{3t-1}(y_t, c_2) > \mu_{2t-1}(y_t)$ then $y_{t-1} > y_t$. Hence to establish claim (a) it is sufficient to show $\mu_{3t-1}(y_t, c_2) > \mu_{2t-1}(y_t)$. Using (3) and (4) it follows

$$\mu_{2t-1}(y_t) = y_t + \beta\mu_{2t}(y_t)$$

and

$$\mu_{3t-1}(y_t, c_2) = y_t - c_2 + \beta(1 - F(y_t)) \cdot \{E(\mu_{2t}(w) \mid w \geq y_t) - \mu_{2t}(y_t)\} + \beta\mu_{2t}(y_t)$$

Manipulating and using the definition of $\mu_{2t}(y_t)$ yields

$$\mu_{3t-1}(y_t, c_2) - \mu_{2t-1}(y_t) = -c_2 + \beta \cdot \int_{y_t}^{\infty} (\psi(w, t + 1) - \psi(y_t, t + 1))f(w)\,dw + \beta(1 - F(y_t))\{E(w \mid w \geq y_t) - y_t\}$$

Using (9) it follows

$$\mu_{3t-1}(y_t, c_2) - \mu_{2t-1}(y_t) = \beta(1 - F(y_t))\{E(w \mid w \geq y_t) - y_t\} > 0$$

Hence it has been established that $y_{t-1} > y_t$ if Strategy A dominates. Claim (b) can be proved by using arguments essentially the same as the above. Finally, claim (c) follows directly from claim (b).

PROOF of Proposition 5:

Substituting (15) and (16) into (17) and manipulating yields

$$V(\tilde{w}, s) = 1 - \lambda(1 - F(\tilde{w}))^s/(1 - F(\tilde{w})) + d(\tilde{w}, s)$$

where $d(w, s) =$

$$m\lambda[F(\tilde{w})]^{t(\tilde{w})-1} - \lambda \sum_{i=1}^{m} [F(y_{N-i})]^{N-i-1}$$

and $m = \min\{s, N - s - 1\}$. Let $h = V(\tilde{w}, s)Q(\tilde{w}, s) - V(\tilde{w}, s - 1)Q(\tilde{w}, s + 1)$. If $h > 0$ for any $s < t(\tilde{w})$, then the major claim of the proposition is established. But, for $s < t(\tilde{w})$,

$$\begin{aligned} h = {} & \lambda[F(\tilde{w})]^{s-1} \\ & \cdot \{1 - F(\tilde{w}) - (1 - [F(\tilde{w})]^{t(\tilde{w})-1})\} \\ & + \lambda[F(\tilde{w})]^{s-1}\{d(\tilde{w}, s)(1 - [F(\tilde{w})]^{t(\tilde{w})-s}) \\ & - d(\tilde{w}, s + 1)F(\tilde{w})(1 - [F(\tilde{w})]^{t(\tilde{w})-s+1})\} \end{aligned}$$

The second term on the right-hand side of the above is positive as $d(\tilde{w}, s) \geq d(\tilde{w}, s + 1)$. Hence, if $1 - F(\tilde{w}) - (1 - [F(\tilde{w})]^{t(\tilde{w})-1}) > 0$ then $h > 0$. This implies that if

$$\text{(A1)} \quad \lambda < (1 - F(\tilde{w}))/(1 - [F(\tilde{w})]^{t(\tilde{w})-1})$$

then $h > 0$. From Proposition 4 and equation (12), we have

$$1 = \sum_{t=2}^{N} \lambda_t = \lambda \sum_{j=2}^{t(\tilde{w})-1} [F(\tilde{w})]^{j-2} + \lambda \sum_{k=t(w)}^{N} [F(y_{k-1})]^{k-2}$$

$$= \frac{(1 - [F(\tilde{w})]^{t(\tilde{w})})}{(1 - F(\tilde{w}))} + \lambda \sum_{j=t(\tilde{w})}^{N} [F(y_{j-1})]^{j-2}$$

Hence

$$\lambda = [1 - F(\tilde{w})] \div \left[1 - [F(\tilde{w})]^{t(\tilde{w})} + (1 - F(\tilde{w})) \sum_{j=t(\tilde{w})}^{N} [F(y_{j-1})]^{j-2}\right]$$

and therefore (A1) is satisfied and the major claim is established. The other claims of the proposition follow directly from this result.

REFERENCES

R. Gronau, "Information and Frictional Unemployment," *Amer. Econ. Rev.*, June 1971, *61*, 290–301.

R. E. Hall, "Turnover in the Labor Market," *Brookings Papers*, Washington 1972, *3*, 709-65.

S. Lippman and J. J. McCall, "The Economics of Job Search: A Survey," *Econ. Inquiry*, June 1976, *14*, 155–89.

J. J. McCall, "Economics of Information and Job Search," *Quart. J. Econ.*, Feb. 1970, *84*, 113-26.

J. P. Mattila, "Job Quitting and Frictional Unemployment," *Amer. Econ. Rev.*, Mar. 1974, *64*, 235-39.

D. T. Mortensen, "Job Search, Duration of Unemployment and the Phillips Curve," *Amer. Econ. Rev.*, Dec. 1970, *60*, 847–62.

———, "On-the-Job Learning About Characteristics," mimeo, Northwestern Univ. 1975.

D. O. Parsons, "Specific Human Capital: An Application to Quit Rates and Layoff Rates," *J. Polit. Econ.*, Nov./Dec. 1972, *80*, 1120-43.

———, "Quit Rates Over Time: A Search and Information Approach," *Amer. Econ. Rev.*, June 1973, *63*, 390–401.

J. Tobin, "Inflation and Unemployment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 1–18.

Related Market Conditions and Interindustrial Mergers: Comment

By MARTIN K. PERRY*

The analysis of M. L. Greenhut and H. Ohta (G-O) in their paper in this *Review* has a number of shortcomings. G-O consider the "maxim" that integration of successive monopolists will lower the price of the final good. They claim this maxim is "mislaid" and that existing demonstrations of it are "sketchy." However, the reasoning behind this maxim is well-established in the literature on vertical integration and is so appealing that no rigor has been demanded.[1] Indeed, the following generalized maxim is clear. Underproduction of the final good is compounded *by imperfect competition* at successive stages, and vertical integration could improve matters.

Define the "upstream" stage as the firms producing the intermediate input and "downstream" stage as the firms producing the final good. G-O's Theorem II suggests a proof of the generalized maxim for the case of an upstream monopolist and an imperfectly competitive downstream stage. However, their proof for even the special case of successive monopolists needlessly employs restrictive demand assumptions. Within their model of successive monopoly, I provide a short and much more general proof which they overlooked.

Secondly, G-O are preoccupied with the problem of whether the price set by the upstream monopolist will be independent of the market structure in the downstream stage. I demonstrate that their result on the invariance of the input price is a special case and is not robust. In addition, it is argued that their preoccupation with this question has no apparent normative justification within their model.

Finally, G-O's analysis of an upstream monopolist integrating forward into a Cournot downstream stage is conceptually incorrect. Their implicit definition of vertical integration is faulty and yields misleading conclusions. I redo the analysis with a proper definition of vertical integration and show that this simple extension merely illustrates the generalized maxim. These three criticisms will be discussed in turn.

I. Successive Monopoly

Greenhut and Ohta pose a model of successive monopolists in which the upstream stage has a constant marginal cost k of producing the intermediate input. In addition, the downstream stage has a fixed-proportions technology. Units of the intermediate input can be defined to correspond to units of the final good, and the marginal cost of producing the final good, excluding the cost of the intermediate input, is a constant c.[2] The inverse final demand function is $f(q)$ where $f' < 0$. The marginal revenue or correspondent function of f is $g(q)$, i.e., $g = f + q \cdot f'$. Similarly, the correspondent function of g is $h(q)$, i.e., $h = g + q \cdot g'$. If

*Research economist, Bell Laboratories. I wish to thank Elizabeth E. Bailey, John C. Panzar, and Lawrence C. Rafsky for helpful comments. This paper represents my own views and not necessarily those of the Bell System.

[1]See the articles by Joseph J. Spengler and by Fritz Machlup and Martha Taber. Machlup and Taber cite other authors. The industrial organization textbook by Frederick M. Scherer also has a discussion of this result on p. 250. In addition, Greenhut and Ohta are mistaken when they say that "... the literature which contends (as we do below) that important benefits stem from vertical integration has been predicated essentially on the analysis of bilateral monopoly" (p. 267). Inefficiencies created by a *single* imperfectly competitive firm can also be eliminated by vertical integration. See Richard Schmalensee, Frederick R. Warren-Boulton, and the author. For a summary of other aspects of vertical integration, see Oliver E. Williamson.

[2]If the production function is $q = \min\{L/\alpha, K/\beta\}$, then redefine the quantity of the intermediate input L to be L/α. If p_K is the price of the other input K, then $c = \beta \cdot p_K$.

disintegrated, the downstream monopolist's inverse derived demand is $g(q) - c$ (equation (8)).[3] The upstream monopolist's marginal revenue is then $h(q) - c$, and its profits are maximized by producing the input quantity q_{II} defined by the first-order condition $h(q_{II}) = c + k$ (equation (9)). Thus, q_{II} is the final output when the monopolists are not integrated. On the other hand, if the monopolists are integrated, the marginal revenue of the joint firm is $g(q)$, and the profit-maximizing final output q_I is defined by the first-order condition $g(q_I) = c + k$ (equation (6)). Since output q_I maximizes *industry* profits, there is clearly an incentive to integrate if $q_{II} \neq q_I$. As a result of the production assumptions on the downstream stage, the integrated outcome q_I is equivalent to that which occurs when the upstream monopolist is not integrated and the downstream industry is competitive.

To prove that integration of the successive monopolists lowers the final price ($f(q_{II}) > f(q_I)$ or $q_{II} < q_I$), G-O assume that 1) $g' < 0$, 2) $h' < 0$, 3) the horizontal lengths of f and g are in a fixed proportion, and 4) the elasticity of final demand is decreasing in price. These assumptions are designed to obtain an intermediate result on the invariance of the input price. For this reason, their proof of lower final prices after integration is needlessly inefficient. I show that assumption 1) is alone sufficient to prove this result.

The assumption that $g' < 0$ means that the marginal revenue function with respect to final demand is downward sloping. Since $f' < 0$, $g' = 2 \cdot f' + q \cdot f'' < 0$ is considered to be the normal case.[4] From the definition of h, $g' < 0$ implies that $g(q) > h(q)$ for all q. In particular, $g(q_{II}) > h(q_{II}) = c + k = g(q_I)$. Clearly, $g' < 0$ and $g(q_{II}) > g(q_I)$ imply that $q_{II} < q_I$. Final prices are lower after integration of the successive monopolists.

II. The Input Price

Greenhut and Ohta's primary purpose for assuming 1) - 4) is to prove that the input price set by the upstream monopolist will be the same whether the downstream stage is competitive or monopolistic. Not only are these assumptions restrictive, but the result is not robust with respect to them. This can be demonstrated by characterizing the general demand functions which yield the invariance of the input price when marginal costs c and k are constant.

When the downstream stage is competitive, the input price is $f(q_I) - c$. When the downstream stage is monopolistic, the input price is $g(q_{II}) - c$. The question is, what functional forms of f imply that $f(q_I) = g(q_{II})$? Since $q_I = g^{-1}(c + k)$ and $q_{II} = h^{-1}(c + k)$, f must be such that $fg^{-1} = gh^{-1}$. The form $f(q) = a - b \cdot q^{\alpha}$ (for appropriate parameters) employed by G-O is such a function. However, despite the fact that $fg^{-1} = gh^{-1}$ for linear and constant elasticity demand curves, this property is not general. If $fg^{-1} > gh^{-1}$, the input price charged to a competitive industry is greater, whereas if $fg^{-1} < gh^{-1}$, the input price charged to a monopolist is greater. There is no reason to expect either of these two cases to be unreasonable or even unlikely. Thus, G-O's result is a special case, and any deviation from the strict condition that $fg^{-1} = gh^{-1}$ yields a different conclusion. Moreover, with decreasing rather than constant returns to the intermediate input, its price is not generally invariant for even linear and constant elasticity demand curves (see Terry G. Foran's results in conjunction with equation (10) of G-O).

Within this model, even a general treatment of the input price has no apparent normative justification. Given an upstream monopolist, having a downstream monopolist (successive monopoly) is less desirable *irrespective of the input price* than having a

[3]This assumes no exercise of monopsony by the downstream monopolist. Such would be the case if there were many other buyers of the input. However, the conditions, particularly the segregation of buyers, which would allow one to ignore these other markets in the analysis (as G-O implicitly do) might make the lack of monopsony a questionable assumption. I make nothing more of this point and employ G-O's model as is.

[4]If $g' > 0$ were allowed over some ranges, there could easily be more than one local profit maximum for the integrated as well as the disintegrated case. I suspect that it can be shown that the global maxima would occur in the same order, i.e., $q_{II} < q_I$. However, such a demonstration is unnecessary for our purpose here.

downstream competitive industry (which is equivalent to having integrated monopolists in this model). For the latter cases, I have shown both that industry profits are greater and that final prices are lower, independent of the input price. Moreover, input price changes cannot affect productive efficiency because the fixed-proportions technology eliminates the possibility of inefficient substitution. Thus, the welfare analysis for this model requires no information about the input price.

III. The Cournot Downstream Stage

Greenhut and Ohta purport to generalize their basic model by allowing a downstream stage of m identical Cournot firms. If $m = 1$, there are successive monopolists, whereas if $m \rightarrow \infty$, the downstream stage becomes competitive. However, I now show that their model improperly generalizes vertical integration.

The inverse derived demand by the Cournot downstream stage for the intermediate input is $f(q) + (q/m) \cdot f'(q) - c$ (equation (14')). An input price $r > k$ is then set by the upstream monopolist so as to induce the input purchases and final output $\hat{q}$ which maximizes its profits. The first-order condition defining $\hat{q}$ is $g(\hat{q}) + (\hat{q}/m) \cdot g'(\hat{q}) = c + k$.[5] When the upstream monopolist now integrates with some of the downstream firms, his input is transferred at marginal cost k to these subsidiaries. *However, in their model the output decisions of the subsidiaries are not coordinated.* Each subsidiary continues to make its output decision in a Cournot fashion with respect to other subsidiaries as well as nonsubsidiaries. The subsidiaries are distinguished only by their lower marginal costs, i.e., $c + k < c + r$. This fact is evident from comparing equations (12) and (18).

The net effect of G-O's implicit definition of vertical integration is that the upstream monopolist relinquishes its monopoly and allows the downstream stage to compete away the monopoly profits. With complete integration, all firms receive the input at marginal cost k, and the final output $\bar{q}$ is defined by $f(\bar{q}) + (\bar{q}/m) \cdot f'(\bar{q}) = c + k$ (equation (18)). If $g' < 0$ (assumption 1)), then final prices are reduced by this integration, i.e., $\hat{q} < \bar{q}$.[6] However, this result does not stem from vertical integration (as G-O assert), but instead from the elimination of the upstream monopoly. Thus, drastically lower prices should not be surprising since even a small number of Cournot firms would compete away the bulk of the original markup over marginal cost k by the disintegrated upstream monopolist.

Greenhut and Ohta suggest that the results of their model are applicable to cases where firms did or desired to integrate forward. However, they failed to notice the strong disincentives for vertical integration as it is improperly defined in their model. We can examine the incentives by comparing *industry* profits before and after integration. The disintegrated profits are $\hat{\pi}(m) = [f(\hat{q}) - c - k] \cdot \hat{q}$, while the integrated profits are $\bar{\pi}(m) = [f(\bar{q}) - c - k] \cdot \bar{q}$. If $\hat{\pi}(m) < \bar{\pi}(m)$, then vertical integration would be profitable. However, if $\hat{\pi}(m) > \bar{\pi}(m)$, then vertical integration would be unprofitable.

Obviously, $\hat{\pi}(1) < \bar{\pi}(1)$ since successive monopoly is eliminated. However, for a somewhat larger number of downstream firms, the previous discussion suggests that industry profits should be reduced by integration. By specifying a functional form for f, we can obtain an idea of how large m must be for the latter case to occur.[7] For f linear, $f(q) = a - b \cdot q$, we find that $\hat{\pi}(2) = \bar{\pi}(2)$ and $\hat{\pi}(m) > \bar{\pi}(m)$ for $m > 2$.[8] For f of constant elasticity, $f(q) = b \cdot q^{-1/\eta}$ $(\eta > 1)$,

[5]At this point, G-O are again sidetracked by the input price question. They show that for $f(q) = a - b \cdot q^{\alpha}$, the monopolist's profit-maximizing input price is invariant with respect to the number of downstream firms m. My previous criticisms again apply.

[6]Since $f' < 0$, $g(q) \leq f(q) + (q/m) \cdot f'(q)$ for $m \geq 1$. Thus, $g(\bar{q}) \leq c + k$. If $g' < 0$, $g(\hat{q}) > c + k$. Therefore, $g(\hat{q}) > g(\bar{q})$ implies that $\hat{q} < \bar{q}$.

[7]These special cases of f are employed to *illustrate* the extent of a known result rather than to *prove* such a result.

[8]Solving for $\hat{q}$ and $\bar{q}$ in terms of m yields

$$\hat{q} = \left[\frac{m \cdot (a - c - k)}{2b \cdot (m + 1)}\right] \quad \text{and} \quad \bar{q} = \left[\frac{m \cdot (a - c - k)}{b \cdot (m + 1)}\right]$$

Substituting into $\hat{\pi}(m) - \bar{\pi}(m)$ and simplifying yields the stated result.

we find that $\hat{\pi}(m) > \bar{\pi}(m)$ for $m \geq 2$ at any demand elasticity $\eta > 1$ (see the Appendix for the proof). This indicates that as few as two downstream firms will make integration unprofitable in the G-O model.

Vertical integration in the G-O model is certainly beneficial since the upstream monopoly is eliminated by definition and final prices considerably reduced. However, since there is typically no incentive for integration defined in this manner, force would be required to achieve their desired result. The upstream monopolist will not choose to freely relinquish its monopoly position. In short, vertical integration is incorrectly specified and the conclusions from the model are unfounded.

We can quickly redo the analysis in a correct manner by allowing the upstream monopolist to coordinate the decisions of its subsidiaries after integration. Thus complete integration results in a single monopolist maximizing industry profits at q_I as before, i.e., $g(q_I) = c + k$. Since the disintegrated output $\hat{q}$ was defined as $g(\hat{q}) + (\hat{q}/m) \cdot g'(\hat{q}) = c + k, g' < 0$ (assumption 1) implies that $g(\hat{q}) > c + k = g(q_I)$. Thus $\hat{q} < q_I$. There is an incentive to integrate (since $\hat{q} \neq q_I$), and final prices are reduced by this properly defined integration (since $\hat{q} < q_I$). The insight is the same as in the case of successive monopolists. Successive stages of imperfect competition compound the underproduction of the final good, even when no substitution possibilities exist. Allowing vertical integration in such cases could result in lower final prices.

APPENDIX

Solving for $\hat{q}$ and $\bar{q}$ in terms of m yields

$$\hat{q} = \left[\frac{c + k}{b \cdot \left(\frac{\eta + 1}{\eta}\right) \cdot (1 - 1/m\eta)}\right]^{-\eta}$$

and

$$\bar{q} = \left[\frac{c + k}{b \cdot (1 - 1/m\eta)}\right]^{-\eta}$$

Substituting into $\hat{\pi}(m) - \bar{\pi}(m)$ and simplifying yields $\hat{\pi}(m) - \bar{\pi}(m) > 0$ for $m > [\eta/(\eta - 1)]^{\eta-1} + 1/\eta - 1$. Thus, $\hat{\pi}(m) - \bar{\pi}(m) > 0$ for all $m \geq 2$ if $[\eta/(\eta - 1)]^{\eta-1} + 1/\eta - 1 < 2$ for $\eta > 1$. Let $s = \eta - 1$. We must then show that $3 - [1 + 1/s]^s > 1/(s + 1)$ for $s > 0$. The term $[1 + 1/s]^s$ approaches 1 as $s \rightarrow 0^+$, approaches e as $s \rightarrow \infty$, and increases monotonically in s. Note that $1/(s + 1)$ decreases monotonically. Now, for $0 < s \leq 1$, $3 - [1 + 1/s]^s \geq 1 > 1/(s + 1)$. Similarly, for $1 < s \leq 3$, $3 - [1 + 1/s]^s > \frac{1}{2} > 1/(s + 1)$. Finally, for $3 < s$, $3 - [1 + 1/s]^s \geq 3 - e > \frac{1}{4} > 1/(s + 1)$. This proves the stated result.

REFERENCES

T. G. Foran, "Market Structure and Derived Demand," *Economica*, Jan. 1976, *43*, 83–87.

M. L. Greenhut and H. Ohta, "Related Market Conditions and Interindustrial Mergers," *Amer. Econ. Rev.*, June 1976, *66*, 267–77.

F. Machlup and M. Taber, "Bilateral Monopoly, Successive Monopoly, and Vertical Integration," *Economica*, May 1960, *27*, 101–19.

M. K. Perry, "The Theory of Vertical Integration by Imperfectly Competitive Firms," unpublished doctoral dissertation, Stanford Univ. 1975.

Frederick M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 442–49.

J. J. Spengler, "Vertical Integration and Antitrust Policy," *J. Polit. Econ.*, Aug. 1950, *58*, 347–52.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, Aug./Sept. 1974, *82*, 783–802.

O. E. Williamson, "The Vertical Integration of Production: Market Failure Considerations," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 112–23.

Related Market Conditions and Interindustrial Mergers: Comment

By JOHN R. HARING AND DAVID L. KASERMAN*

In a recent issue of this *Review*, M. L. Greenhut and H. Ohta (hereafter G-O) claim to have resurrected a *Mislaid Maxim* which they consider "... to be especially vital in this period of energy shortage and attacks on big business" (p. 267). The maxim is that important benefits stem from vertical integration. In particular, their results indicate that "... the benefits stemming from vertical merger of successive monopolies dovetail with the effects of having perfect competition on the higher stage of production rather than monopoly" (p. 267). G-O express hope that, given their identification of this result and the generality of their proofs, this maxim will no longer strike anyone as surprising.

Greenhut and Ohta deduce two theorems:

> ... (I) *the price charged by the monopolistic supplier of an input is not affected by the market structure in the market for the final good, be it perfectly competitive, monopolistic, etc., and* (II) *merger or collusion between the input supplier and the final good producer brings about lower prices, greater output and sales, and greater profits to the merged or colluding firms—a welfare gain.* [p. 267]

Self-advertisement notwithstanding, G-O's "general" proofs depend critically upon a restrictive set of assumptions. Their failure to appreciate the importance and circumscriptive nature of these assumptions leads them to infer sweeping policy prescriptions when none appear warranted. A more thorough review of the literature would have revealed that the *maxim* in question, has, in fact, not even been *mislaid*. It has merely been consigned to that long shelf with other truths of recognized validity but limited relevance.

Conditions necessary for the proof of Theorem I are: 1) that final product inverse demand functions $[p = f(q)]$ yield correspondents $[g(q) = f(q) + f'q]$ such that the horizontal lengths to the two curves remain in fixed proportion; 2) that the final output be obtained from a fixed-proportion production function; and 3) that the monopolized intermediate product be subject to constant cost production.

The second assumption (fixed-proportion downstream production) is a necessary condition for the input monopolist's average revenue function to be obtained directly from the final product inverse-demand function for a competitively structured final good industry, or from its correspondent for a monopolistic downstream industry. As is well-known, when substitution possibilities are present, the derived demand for a given input is in general a function of final output price, production parameters, and the employment of all other factors. Hence, a unique mapping from final product demand to intermediate product demand does not, in general, exist.[1] As shown by John Vernon and Daniel Graham, Richard

*Staff economists, Bureau of Economics, Federal Trade Commission. The views expressed herein are not intended and should not be construed as representative of an official Commission policy. We benefited from discussions with P. David Qualls, John J. Siegfried, and F. M. Scherer.

[1]Albert Rees, p. 86, notes that in the long run, with constant costs and linear demand curves, the elasticity of derived demand will not change as a competitive industry becomes monopolized. Charles Maurice and C. E. Ferguson have also examined the effect of monopoly upon derived demand. In their long-run analysis, the market structure term is the difference between the elasticity of marginal revenue and the elasticity of marginal cost. They conclude that this term is related to the elasticity of commodity demand, but that the relation is tenuous and cannot be stated explicitly in meaningful economic terms. More recently, Terry Foran has analyzed the effect of monopoly upon derived demand under short-run industry conditions. He concludes that the manner in which market structure affects the elasticity of derived demand at specific input price levels must be viewed as an empirical question.

Schmalensee, George Hay, and Frederick Warren-Boulton, substitutability in the production of the final good prevents the input monopolist from extracting full monopoly rents and provides an incentive for forward integration not present under fixed proportions. With a competitive downstream industry, the welfare effects of such integration cannot be determined a priori, but depend upon the demand and production conditions in the industry. In essence, forward integration in this situation leads to increased efficiency of resource utilization within the integrated firm. This increased efficiency may, however, be partially or completely offset by: a) increased distortions in input utilization by nonintegrated firms as the profit-maximizing price of the monopolized intermediate good is altered by a reduction in derived demand as the proportion of final output produced by the integrated firm increases (Schmalensee); and b) final product price and output distortions that derive from the input monopolist's acquisition of downstream monopoly power (Warren-Boulton).[2]

The necessity of the third assumption (constant cost production of the monopolized input) can easily be demonstrated by reproducing G-O's Figure 1 with input production subject to increasing marginal cost. As shown in our Figure 1, allowing k (the marginal cost of the monopolized input) to be an increasing function of q (valid under fixed proportions) results in a divergence in the profit-maximizing price of the intermediate good with alternative downstream market structures. In the graph, p_1 will obtain with a competitive output market and p_2 will result with a downstream monopoly, thus refuting Theorem I.

With regard to Theorem II, the crucial assumption (which mysteriously disappears when G-O turn to the policy implications of their results) is that monopoly exists at each of two successive stages of production, but

FIGURE 1

that the downstream monopolist exercises no monopsony power. Given this assumption, an unambiguous gain in net social welfare can be reaped by vertical control either through collusion or integration under general production and demand conditions. Wesley Liebeler refers to this case as the "repeated marginalization of successive monopolies." It is referred to by Frederick Scherer as one of "myopic chain monopoly," and its validity is proven in a footnote, p. 243, under assumptions of constant costs, fixed proportions, and linear demand. Other sources which demonstrate the result, and which along with Liebeler and Scherer are not cited by G-O, include Joseph Spengler, Eugene Singer, and Hay.

The introduction of monopsony power on the part of the downstream monopolist results in the bilateral monopoly situation which renders input price indeterminate.[3] But, if bilateral monopolists agree on both

[2]As noted by Warren-Boulton, p. 796, the indeterminacy of the welfare effect of downstream monopolization is suggested by the theory of the second best since this is tantamount to the acquisition of a second monopoly in a world that already suffers from an existing monopoly distortion.

[3]The existence of more than one downstream monopolist, which is required if the bilateral monopoly situation is to be avoided, necessitates the selling of final output in multiple markets—either attributable to geographical dispersion of monopolists of a single product, or employment of the monopolized input in the production of more than one final good. In either case, vertical cooperation may be difficult because of information and control problems that emanate from such diversity. See Oliver Williamson.

the price *and* quantity of the intermediate good exchanged, the outcome in the final product market will be identical to that obtained under integrated monopoly since the maximization of joint profits is necessary if the bargaining parties are to remain on the contract curve.[4] Allocation of these profits is then determined by the contracted price of the intermediate good.

The identical nature of the welfare results that stem from the integrated successive monopoly and the bilateral monopoly cases is not, however, the cause of Machlup and Taber's (nor others') reluctance to infer broad policy implications from the successive monopoly argument. Nor does such hesitancy stem from any lack of proven theorems in the literature. Rather, the general unwillingness to draw such conclusions (which G-O mistake for the purloining of an important proposition) appears to reflect a recognition of the possibility that vertical integration might serve to solidify or expand the monopoly power that provided the original catalyst for vertical control.[5]

[4]See A. L. Bowley, James Morgan, William Fellner, and Fritz Machlup and Martha Taber.

[5]This possibility is noted by Alfred Marshall (see the quote on p. 117 of Machlup and Taber). A similar potential tradeoff between short-run static efficiency and long-run competitive markets is noted by Kenneth Arrow, p. 181, where the incentive to integrate vertically derives from an assumed asymmetry in information between upstream and downstream firms.

REFERENCES

K. J. Arrow, "Vertical Integration and Communication," *Bell J. Econ.*, Fall 1975, *6*, 173–83.

A. L. Bowley, "Bilateral Monopoly," *Econ. J.*, Dec. 1928, *38*, 651–59.

W. Fellner, "Prices and Wages Under Bilateral Monopoly," *Quart. J. Econ.*, Aug. 1947, *61*, 503–09.

T. G. Foran, "Market Structure and Derived Demand," *Economica*, Feb. 1976, *43*, 83–87.

M. L. Greenhut and H. Ohta, "Related Market Conditions and Interindustrial Mergers," *Amer. Econ. Rev.*, June 1976, *66*, 267–77.

G. A. Hay, "An Economic Analysis of Vertical Integration," *Ind. Org. Rev.*, 1973, No. 3, *1*, 188–98.

W. J. Liebeler, "Toward a Consumer's Antitrust Law: The Federal Trade Commission and Vertical Mergers in the Cement Industry," *UCLA Law Rev.*, June 1968, *15*, 1153–202.

F. Machlup and M. Taber, "Bilateral Monopoly, Successive Monopoly, and Vertical Integration," *Economica*, May 1960, *27*, 101–17.

Alfred Marshall, *Principles of Economics*, London 1920.

C. S. Maurice and C. E. Ferguson, "Factor Demand Elasticity Under Monopoly and Monopsony," *Economica*, May 1973, *40*, 180–86.

J. N. Morgan, "Bilateral Monopoly and the Competitive Output," *Quart. J. Econ.*, Aug. 1949, *63*, 371–91.

Albert Rees, *The Economics of Trade Unions*, Chicago 1962.

Frederick M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

R. Schmalensee, "A Note on the Theory of Vertical Integration," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 442–49.

Eugene M. Singer, *Antitrust Economics: Selected Legal Cases and Economic Models*, Englewood Cliffs 1968.

J. J. Spengler, "Vertical Integration and Antitrust Policy," *J. Polit. Econ.*, Aug. 1950, *58*, 347–52.

J. M. Vernon and D. A. Graham, "Profitability of Monopolization by Vertical Integration," *J. Polit. Econ.*, Sept./Oct. 1971, *79*, 924–25.

F. R. Warren-Boulton, "Vertical Control with Variable Proportions," *J. Polit. Econ.*, July/Aug. 1974, *82*, 783–802.

O. E. Williamson, "The Vertical Integration of Production: Market Failure Considerations," *Amer. Econ. Rev. Proc.*, May 1971, *61*, 112–23.

S. Y. Wu, "The Effects of Vertical Integration on Price and Output," *Western Econ. J.*, 1964, No. 2, *2*, 117–33.

Related Market Conditions and Interindustrial Mergers: Reply

By M. L. Greenhut and H. Ohta*

Martin Perry happily accepted our results but reworked them in a way he considered more efficient. We appreciate his proving the results given in our paper, but do not agree that his technique (based on our paper) is the better one. Before buttressing this claim, let us observe that if he did not believe rigor is needed to establish the subject Mislaid Maxim, he should not have taken the time to recast our model in what he considers a more efficient form. Apparently he really believes the Mislaid Maxim does require the most penetrating analysis possible.

Let us save space by considering his Sections I and II together. We agree $g' < 0$ is sufficient for our Theorem II.[1] We do *not agree*, however, that Theorem I is unimportant, and hence that simply specifying $g' < 0$ is alone required. More specifically, Perry's statement that Theorem I is special implies its unimportance. But how many readers of the original paper or of these notes alone would agree with him that demonstrating the *invariance of a monopolist's* input supply price with respect to the market structure in the next stage, whether the final product sells at the perfectly competitive or a simple monopoly price, is per se unimportant? Rather in our view, this theorem is interesting, not intuitively obvious, and for reasons to be given below, very important.

Perry's next critique that the relations $fg^{-1} \gtrless gh^{-1}$ may hold is surprising in light of his emphasis on rigor "cum" simplicity, though let us say at the outset that the possibility of these relations was certainly not ignored when we specified the general form of f that establishes the equality.[2] Most fundamentally, Perry cannot simply assert, *as he did*, the importance of the unequal relations and hang us to the cross because we stressed $f[g^{-1}(c + k)] = g[h^{-1} \cdot (c + k)]$. Not only did he fail to demonstrate the superior relevance (and *robust* qualities) of the form he proposed, but even more surprisingly he completely misinterpreted the form he rejected, as is manifest by his contention that it provides either linear or constantly elastic demand curves when, in fact, we also demonstrated its relevance to concave and convex demand curves.

His next objection referring to decreasing returns violates an assumption we used in the part of the paper he is critiquing. Not only does he take us out of context by the subject remark, but he ignores the fact that later on in the paper we did change our assumption, in effect indicating the generality or, should we say, inclusiveness of our Theorem II *beyond its relation* to Theorem I. In sum, Perry's Sections I and II establish *on the basis of our model*, albeit in his restricted way, the Mislaid Maxim, Theorem II. We trust that further consideration of it, and also Theorem I, will take place in the years to come.

Perry's only (possible) real critique—other than for his references to what we should or should not have included in our paper—could therefore be contained in his Section III. But here *again* he took us out of context, actually in a more objectionable way than he did in the statement referred to

*Professor of economics, Texas A&M University, and associate professor of economics, Aoyama Gakuin University and University of Houston (adjunct), respectively.

[1] It is certainly the case that our second theorem alone can be deduced from a much less restrictive set of assumptions than that required to deduce both theorems simultaneously. Not only may the form of the demand function be relaxed, but the form of the MC curve can also be relaxed if Theorem II alone were deduced, as we ourselves demonstrated in Section III.

[2] Incidentally, fg^{-1} and gh^{-1} are misleading notations. They could mean $f \times g^{-1}$ and $g \times h^{-1}$, respectively, which are not what Perry means; what he really wants are $f(g^{-1}(c + k)) \gtrless g(h^{-1}(c + k))$.

in the paragraph above. Our evaluation of price effects of merger by a monopolist supplier of final market oligopolists was simply that alone. *Whether or not the firm would merge with just one or two or many downstream firms is basically non sequitur.* But to go with Perry's objective rather than ours, recall his argument that the upstream monopolist in our model allows the downstream stage "to compete away the monopoly profits" after the vertical integration has taken place. This argument is simply invalid. The correct proposition is that the downstream firms merging with the upstream monopolists would compete profits away *from all of the firms that may still remain independent.* What if no independent firms remain, one might ask as does Perry? He raises this query, in effect, as an extreme case since complete vertical integration subject to Cournot horizontal competition may not prove to be profitable to every participant (i.e., it may be profitable to some but not to all participants insofar as industry profits happen to be decreased under the circumstances). This particular possibility does not generate any problem, of course, for - if it happened to be the case—vertical integration would simply proceed in part, as the last firms willing to integrate would not receive an attractive contract offer from the upstream monopolist. Complete vertical integration under these circumstances would require horizontal merger as well for profit incentives. *In any case,* output is increased after vertical integration, be it complete or incomplete, and the conclusion we set forth in our paper remains invariant. Note further that Perry's argument that *industry profits* may be reduced after vertical integration has proceeded to its limit fails to distinguish individual firm incentives from industry profits, and only individual firm incentives count.[3]

May we conclude our reaction to Perry's observations by noting that our demonstration and findings dovetailed fully with the *subject matter we proposed* we would consider. His own analyses *in support of our findings* and his extensions in terms of his more restricted subject matter are reassuring.

The paper by John Haring and David Kaserman (henceforth H-K) argues in perhaps a more confined way than Perry that our assumptions are restrictive and the *Maxim unimportant.* A quick perusal of their critique compared with Perry's might suggest that their claim could be meritorious. For the reasons recorded below, we appreciate their observations, albeit as with Perry *rejecting all of them completely.* Consider the following:

We *did claim* that the Mislaid Maxim is "especially vital in this period of energy shortage and attacks on big business," *and we repeat that claim here.* Haring-Kaserman objected to it by contending that we used overly restrictive assumptions. They cited several economists who worked with models involving variable proportions rather than fixed proportions. Apparently in the H-K view, if the number of economists who use a different assumption than ours happens to be large, our assumption must be the restricted one.[4] We shall answer this particular charge by simply noting that diverse petroleum company representatives have advised us that the quantity of refined products per barrel of crude is essentially invariant to the alternative production processes that may be employed. Fixed proportions thus apply, at least in this part of the energy industry. Moreover, it is clear that besides the refining of crude, the ship-

[3]It is only along the lines of individual incentives for vertical integration that our Cournot assumption could (or should) be questioned. Indeed, if it were not for fear of antitrust prosecution, firm(s) which may collude vertically with input supplier(s) could readily identify the reaction functions of noncolluded firms and lead them accordingly. Then a Stackelberg leader-follower type of behavioral reaction would be required to replace our Cournot reaction. Our welfare conclusion on vertical integration could, in turn, be shown to remain unchanged again, which demonstration would be based on the more realistic condition of imperfect competition in all stages of production. A generalized analysis of vertical integration along this line must be reserved, however, to other writings.

[4]It has been argued by Robert Basmann that there is no sense in which assumptions are too restrictive or unrealistic; but let us accept H-K's words without methodological quibble.

ment of oil from the well to the refinery and the shipment again from the refinery to the dealer (each involving successive stages of production) is in fixed proportions. How much empirical relevance must one demonstrate before the related maxims can be considered important, and the underlying assumptions not considered narrow or restrictive?

The subject writers objected to our failure to include analyses of monopsony and bilateral monopoly in our paper. However, why should a downstream firm necessarily be a monopsonist? And frankly, in the world of spatial oligopoly and oligopsony, with overlapping market and supply areas, monopsony or bilateral monopoly would appear to be very unlikely market types. Let H-K deal with the narrow or restrictive markets, if in fact they really are interested in such markets.

Haring-Kaserman finally argue that they refuted our Theorem I by considering increasing marginal cost. They did this by substituting increasing marginal cost in place of the constant marginal cost that we assumed in Sections I and II of our paper. But which here is the more restrictive? Rather often we theoreticians are advised that real world firms produce at constant marginal costs over a wide range of output.[5] If, *nevertheless*, we do accept *their* increasing or decreasing marginal cost assumption, our Theorem I would indeed be affected, *and we never claimed it would not be!*[6] But we did not, of course, consider generality of assumptions for the sake of generality to be warranted in our paper, especially since, to repeat, constant *MC* and fixed input proportions appear to be *the relevant conditions in the energy and distributive industries to which we directed our analysis*, and hence Theorem I *as stated* applied.[7]

Our position allows only one possible point of agreement between Perry, H-K, and us. And happily we *do agree*, although it is really with only an implicit idea of theirs: we agree the relevance of economic theory and the maxims which stem therefrom would be advanced if economic theory is so cast that it would shed light on important economic situations.

[5]Constant marginal cost applies to petroleum refineries up to around 90 percent of capacity.

[6]Theorem I is readily modifiable to include upward or downward sloping *MC* curves. Either our own Figure 1 or H-K's Figure 1 clearly implies that if the *MC* curve slopes upward (or downward), the input price applicable to the downstream industry would be higher (or lower) under competitive conditions than under conditions of monopoly. In no way is the insight or predictive power of our model restricted to a constant *MC* assumption; only Theorem I *as stated* requires constant *MC*.

[7]Incidentally, the impact of nonconstant marginal cost is not unfavorable to the Mislaid Maxim Theorem II.

REFERENCES

R. L. Basmann, "Modern Logic and the Suppositious Weakness of the Empirical Foundations of Economic Science," *Z. Volkswirt. und Stat.*, Heft 2 1975, *2*, 153–76.

M. L. Greenhut and H. Ohta, "Related Market Conditions and Interindustrial Mergers," *Amer. Econ. Rev.*, June 1976, *66*, 267–77.

J. R. Haring and D. L. Kaserman, "Related Market Conditions and Interindustrial Mergers: Comment," *Amer. Econ. Rev.*, Mar. 1978, *68*, 225–27.

M. K. Perry, "Related Market Conditions and Interindustrial Mergers: Comment," *Amer. Econ. Rev.*, Mar. 1978, *68*, 221–24.

Input Choices and Uncertain Demand: Comment

By WOLFGANG MAYER*

In his article in this *Review*, Duncan Holthausen determines the optimal input choice of perfectly and imperfectly competitive firms under uncertainty. He bases his model on the assumptions that capital is an *ex ante* control, whereas labor is an *ex post* control. Stated differently, he assumes that the capital-input choice must be made before the uncertainty about demand for the final commodity is resolved, whereas labor can be chosen after the demand relations have become known with certainty. Holthausen concludes that for both the competitive and quantity-setting imperfectly competitive firm the cost-minimization criterion is established by the equality of the ratio of input prices with the marginal rate of substitution in production. This relationship is expressed in his equation (12).

My aim in this comment is to show that Holthausen's conclusions do not logically follow from his assumptions. If labor is an *ex post* control, and therefore adjustable after the uncertainty element has disappeared, it is generally not true that the cost-minimizing input combination is attained where the marginal rate of substitution in production is equal to the ratio of input prices. I will also demonstrate that Holthausen reaches his conclusion because his model, in contradiction to the underlying assumptions, implies that both capital *and* labor are treated as *ex ante* controls.

In order to economize on space, Holthausen's notation is employed without repeating its meaning, and my argument is restricted to the case of the perfectly competitive firm.

The assumption of capital being an *ex ante* control while labor is an *ex post* control implies that the competitive entrepreneur is faced with a two-stage decision process. The capital-stock decision is made when price uncertainty is prevailing. On the other hand, the labor-input decision, and therefore also the final output decision, is made at a second stage when the commodity price has become known. At this later stage the competitive firm will hire labor up to the point where

$$w = pQ_L \tag{1}$$

where $Q_L = \partial Q/\partial L$. Given the usual properties of a twice-differentiable, strictly quasi-concave production function $Q = Q(L,K)$, one can solve (1) for the optimal amount of labor, L^*:

$$L^* = L(p,w,K) \tag{2}$$

At this second decision stage, all the independent variables in (2) are known with certainty. Of course, K had been decided upon at an earlier period, when price (p) was still a random parameter.

The capital stock is chosen before p is known with certainty. The entrepreneur's objective is to maximize expected utility from profits, taking into consideration the *ex post* adjustment possibilities that exist with respect to labor.[1] This decision process is described by

$$\max_K E[U(\pi)] \tag{3}$$

where $\pi = [pQ(L,K) - wL - icK]$ and where the labor input in the profit function is determined in equation (2). The first-order condition for a maximum is given by

*Associate professor of economics, University of Cincinnati.

[1] An uncertainty model in which some input decisions have to be made *ex ante* while others can be made *ex post* has been developed by Stephen Turnovsky. He assumes that the firm makes its initial production decision when the selling price is still unknown, but that later adjustments in output are still possible once the actual price has become known. In this context, Turnovsky introduces the idea of a two-stage decision process (pp. 395–96).

(4)
$$E\{U'[(pQ_L - w)(\partial L^*/\partial K) + pQ_K - ic]\} = 0$$

which, after substitution of (1), reduces to

$$(5) \qquad E[U'(pQ_K - ic)] = 0$$

Provided labor is an *ex post* control while capital is an *ex ante* control, equations (1) and (5) represent the conditions for the optimal choice of inputs for the competitive firm. The two equations can then be combined and expressed in the form

(6)
$$\frac{ic}{w} = \frac{E(U'p)Q_K}{E(U')pQ_L} = \frac{Q_K}{Q_L}\left[\frac{\mu}{p} + \frac{cov\,(U',p)}{pE(U')}\right]$$

where $\mu = E(p)$. In writing (6) I utilized the fact that Q_L and Q_K are nonrandom technical relations and that w and ic are exogenously given for the competitive firm. Equation (6) states that the ratio of input prices equals the product of the marginal rate of substitution and of a term which generally will be different from one. This term, stated in the bracket, consists of two components: while $cov\,(U',p)/E(U')p$ is always negative, the value of μ/p may be greater than, equal to, or less than one.

This leads to the conclusion that, under the assumptions made by Holthausen, his proposition concerning the optimal input choice of the competitive firm as expressed in his equation (12) is incorrect.

The reason why Holthausen reaches a different conclusion is quite obvious. The actual specification of his model and the results he derives from it are not consistent with his initial assumptions. In making his assumptions, labor is made explicitly an *ex post* control. But when Holtausen specifies the labor requirement function in his equation (6) and then maximizes expected utility with respect to capital *and* output in his equation (7), he in fact treats labor as an *ex ante* control. Given the production function $Q = Q(L,K)$, whenever both capital and output are *ex ante* choices it is inevitably implied that labor is chosen *ex ante* as well. This specification, however, runs counter to the initial assumption of labor being variable after the uncertainty had been resolved. The crucial point is that output in equation (7) of Holthausen's paper must not be treated as an *ex ante* decision variable. As long as labor can be adjusted *ex post*, output can be adjusted *ex post* as well. The situation is exactly that of the common short-run competitive production model, where labor is the variable input. For a given capital stock—in the present case chosen under uncertainty at an earlier point in time—the optimal output level is determined where short-run profits are a maximum. Hence, not only labor input, but output itself is variable in the short run. Summing up my argument, Holthausen is in error when, in his equation (7), he makes output a decision variable. At this stage of input choice only capital is a decision variable.

I stated before that Holthausen's equations (6) and (7) imply that, contrary to his assumptions, both capital and labor are *ex ante* controls. It is quite correct that in this case the competitive firm's optimal input combination will be chosen in such a way that the marginal rate of substitution in production equals the ratio of input prices. This conclusion has also been reached by Raveendra Batra and Aman Ullah, p. 548, and more recently been confirmed by Richard Hartman, p. 1290.

REFERENCES

R. N. Batra and A. Ullah, "Competitive Firm and the Theory of Input Demand under Price Uncertainty," *J. Polit. Econ.*, May/June 1974, *82*, 537–48.

R. Hartman, "Competitive Firm and the Theory of Input Demand under Price Uncertainty: Comment," *J. Polit. Econ.*, Dec. 1975, *83*, 1289–90.

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

S. J. Turnovsky, "Production Flexibility, Price Uncertainty and the Behavior of the Competitive Firm," *Int. Econ. Rev.*, June 1973, *14*, 395–413.

Input Choices and Uncertain Demand: Reply

By DUNCAN M. HOLTHAUSEN*

In his comment, Wolfgang Mayer claims that some conclusions in my earlier paper in this *Review* do not logically follow from my assumptions. I think it would be more accurate to say that there is some ambiguity in my assumptions. Be that as it may, Mayer has interpreted the assumptions differently than I did, and my purpose in writing this reply is to show the shortcomings of his interpretation.

I had assumed that the firm faced a random demand curve and had to select the level of capital stock before the demand uncertainty was resolved. Thus, capital is an *ex ante* control. Labor, on the other hand, was assumed to be more flexible and could be adjusted after the uncertainty was resolved. Labor is therefore an *ex post* control. The problem arises in the cases of the competitive firm and the quantity-setting imperfect competitor. I assumed that these firms had to set output before the uncertainty was resolved. Hence, output also is an *ex ante* control. Now if capital is an *ex ante* control, we cannot have output chosen *ex ante* and labor chosen *ex post* as Mayer has pointed out. Output and labor must both be chosen *ex ante* or they must both be chosen *ex post*. At this point, one must decide which assumption fits best. In writing my paper I implicitly assumed that quantity-setting firms had to choose output *ex ante*, and that labor was thus determined *ex ante* also. This does not contradict the original assumption that labor can be adjusted *ex post*, it just means that quantity-setting firms cannot avail themselves of labor's flexibility. Since they must make both capital and output decisions *ex ante*, they are locked into a particular amount of labor *ex ante*.

*Associate professor, North Carolina State University.

Mayer, on the other hand, has assumed that quantity-setting firms may adjust output after the demand uncertainty is resolved. Thus output becomes an *ex post* control permitting labor to remain an *ex post* control. The difficulty with this assumption is that the firm somehow discovers what price will prevail before it makes its output decision.[1]

This is quite unreasonable for the quantity-setting imperfect competitor. Since it faces a downward-sloping demand curve, the price that eventually prevails must depend upon the level of output chosen by the firm as well as the random element of demand.

In the case of the perfectly competitive firm analyzed by Mayer in his comment, the assumption that output can be chosen *ex post* ignores some important general equilibrium considerations. First, if all firms behave this way, we must question how a market price can be determined before any firms have made output decisions. Second, even if an initial market price is determined, it ought to change after all firms have made their *ex post* labor and output decisions. In this case, what occurs is a series of output adjustments followed by market price changes followed by further output adjustments, and so on. Mayer's analysis is correct only in the case where one competitive firm can make *ex post* output decisions while all others must make them *ex ante*. But this is not a very interesting case. In perfect competition, we should not expect one firm to have such production flexibility while it is denied to all others.

What Mayer has done in his comment is

[1]Mayer states after equation (2) that price, the wage rate, and capital stock are all known before the firm chooses its labor input and hence its output.

to suggest an interesting alternative to the implicit assumption made in my paper. It would be interesting to work out the full general equilibrium model to determine the ultimate effect on input choices. As matters stand, however, Mayer's results must be viewed as tentative until the full model is developed.

REFERENCES

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

W. Mayer, "Input Choices and Uncertain Demand: Comment," *Amer. Econ. Rev.*, Mar. 1978, *68*, 231–32.

Excess Burden: The Corner Case vs. Ballentine and McLure

By JOHN G. HEAD AND CARL S. SHOUP*

One of the most important and best known propositions in the traditional analysis of excise taxes is that the excess burden of an excise varies directly with the degree of substitutability in consumption between the taxed good and other goods. This conclusion seems, however, to conflict with the intuition, common among students, that, even abstracting from revenue differences, the consumer should suffer less of a burden in switching his consumption from the taxed good the better the substitutes that are available. In a previous paper published in this *Review* we tried to show that this intuition may have some limited validity in cases of corner solutions which have been somewhat neglected in the excess burden literature.

Following Arnold Harberger's well-known diagrammatic analysis, and assuming linear demand schedules and constant costs, we showed that, for a given tax per unit, as substitutability increases from zero and the slope of the income-compensated demand schedule steadily diminishes, excess burden increases from zero to a maximum which is reached where the tax is just sufficient to choke off completely consumption of the taxed product. Beyond this point further increases in substitutability produce corner solutions and excess burden steadily diminishes, ultimately to zero where perfect substitutes are available. The general proposition that excess burden varies directly with substitutability in consumption therefore requires qualification when account is taken of these "corner cases."

In their recent note Gregory Ballentine and Charles McLure (hereafter called B-M) take issue with this analysis, arguing that our results depend crucially on the assumption of constant costs. According to them, when full account is taken of the loss of producer's as well as consumer's surplus, the traditional result that the measure of excess burden of an excise tax is positively related to substitutability in demand is reestablished. They purport to demonstrate this conclusion, first with the aid of diagrammatic analysis for the special case of perfect substitutability, and then with a general equilibrium mathematical model for intermediate cases. It seems clear, however, that their analysis completely ignores the corner cases which were the subject of our original note.

In the case of perfect substitutability, and assuming increasing opportunity costs, they show in their Figure 1 the familiar case in which the excise tax moves the economy from an initial Pareto optimal tangency equilibrium at *A* to an inefficient tangency equilibrium at *B* on the transformation curve. In contrast to our constant cost case, under increasing costs some excess burden results due to the loss of producer's surplus, even though substitutability is perfect. But the inefficient regular tangency equilibrium at *B* in the B-M diagram is clearly unrelated to the corner solutions which we were concerned to analyze. This confusion in the B-M analysis evidently results from their mistaken identification of our "corner cases" with the case of perfect substitutability in consumption.

Their diagram does, however, serve to draw attention implicitly to the significant possibility (which cannot arise in our constant cost case) that, unless the tax rate is sufficiently high, increasing substitutability in demand may not result in corner solutions. In terms of our original discussion, the smaller the supply elasticity, the higher the minimum tax rate required to guarantee the eventual emergence of corner solutions as substitutability in demand increases. As-

*Professor of economics, Monash University, and professor emeritus of economics, Columbia University, respectively.

suming the tax rate is sufficiently high in this sense, our argument that excess burden declines as substitutability increases beyond the point at which consumption of the taxed good is just reduced to zero, can, of course, be illustrated equally well in the B-M diagram by introducing social indifference curves reflecting differing degrees of substitutability and intersecting the transformation curve on the vertical axis. It is certainly true, as their argument shows, that under increasing costs excess burden will not decline to zero (as in our constant cost example), but it will nevertheless decline, specifically to the level of producer's surplus in the initial pretax equilibrium.

The same confusion is evident in their otherwise elegant and useful application of a Harberger-style general equilibrium mathematical model to the "intermediate cases" of less-than-perfect substitutability. The expression which they derive for the excess burden is valid only for a comparison of regular tangency equilibria or intersections of demand and supply schedules. It is clearly not applicable in cases of corner solutions. This is easy to see from their equation (2) which would imply that a doubling, say, of a tax which is already sufficient to choke off demand would double the excess burden, whereas excess burden would actually remain unaltered.

Briefly then, and apart from the important proviso that the tax rate must be sufficiently high, the introduction of producer's surplus makes no essential difference to our argument. In fact the degree of substitutability in production plays very much the same role as substitutability in consumption. Thus, in terms of our original diagram, it is easy to see that as substitutability in production increases from zero (with a vertical supply schedule), excess burden steadily increases up to a maximum where, for a given tax per unit and given demand schedule, consumption of the taxed product is just reduced to zero. Once this point is reached, however, further increases in the supply elasticity will reduce excess burden.

As recognized in our original note, it is of course true that the case in which market demand is completely choked off is of very limited practical interest. Even the most severe sumptuary excises do not normally aim to reduce consumption, and hence revenue, to zero. Practical examples might, however, be found in the related area of fines and penalties. It is also of some interest to notice that the same line of argument applies to all policies which operate via direct restrictions on output, such as quotas and regulations. For a given arbitrary reduction in quantity, excess burden is smaller the greater the degree of substitutability in consumption and production.

Moreover it is quite likely that at least some consumers will be driven out of the market by an excise tax even where total output remains positive. In such cases it is interesting to notice that measures of excess burden, such as those offered by Harberger or B-M, based on the elasticity of market demand at the original equilibrium would tend to be upward biased as they overstate the welfare loss for those consumers for whom corner solutions apply. When the number of such consumers is significant these traditional measures may therefore be inadequate. And in terms of our original paradox it is also true that for these consumers a higher degree of substitutability will reduce excess burden.

REFERENCES

J. G. Ballentine and C. E. McLure, Jr., "Excess Burden: The Corner Case in General Equilibrium," *Amer. Econ. Rev.*, Dec. 1976, *66*, 944–46.

A. C. Harberger, "Taxation, Resource Allocation, and Welfare," in *The Role of Direct and Indirect Taxes in the Federal Revenue System*, Nat. Bur. Econ. Res. and Brookings Inst., conference report, Princeton 1964.

J. G. Head and C. S. Shoup, "Excess Burden: The Corner Case," *Amer. Econ. Rev.*, Mar. 1969, *59*, 181–82.

NOTES

The ninety-first annual meeting of the American Economic Association will be held in Chicago, Illinois, August 29-31, 1978.

The 1978 Employment Registry and Center will be held December 28-30 at the Conrad Hilton Hotel, Chicago, Illinois. Please note that there will be no employment service at the August meetings. Complete details will be published in the June 1978 issue of this *Review*.

Economists who are *strongly* oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Specifically, economists may be eligible if (a) they deal with the history of economic thought or economic history, and (b) if their approach is qualitative and descriptive rather than quantitative and statistical. Conferences dealing with the establishment of social policy or legislation are ineligible. The deadlines for applications to be received in the *ACLS* office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the *ACLS* (Attention: Travel Grant Program), 345 East 46th St, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

American scholars and other professionals interested in university lecturing or research abroad are invited to register with the *CIES*. Registrants receive announcement of the competition in March or April, for appointments to begin twelve to eighteen months later. The general competition for Australia, New Zealand, and Latin American countries closes June 1, and for other countries, July 1. The general composition of the program involving more than 70 countries is expected to be similar to that of recent years. Registration for personal copies of the announcement is now open; forms are available from the Council for International Exchange of Scholars, Suite 300, Eleven Dupont Circle, Washington, D.C. 20036.

On April 28-30, 1978, the Northeast Peace Science Society will hold its fifth annual convention at State University of New York-Binghamton. Interested parties may contact Jack Duffy, School of Management, SUNY-Binghamton, Binghamton, NY 13901.

Omicron Delta Epsilon, the International Honor Society in Economics announces the co-winners of the eighth annual Irving Fisher Monograph Award: Gerard R. Butters, Ph.D., University of Chicago, "Equilibrium Price Distributions and the Economics of Information," and Sanford J. Grossman, Ph.D., University of Chicago, "Essays on Rational Expectations, the Informational Role of Futures Markets, and Equilibrium Bayesian Experimentation." In addition, Lawrence Herbst, Ph.D., University of Pennsylvania, "Interregional Commodity Trade from the North to the South," received Honorable Mention.

The Committee on Social Implications of Technology of the Communications Society (Institute of Electrical and Electronics Engineers) is soliciting submission of papers in this general area for publication either in the IEEE Transactions on Communications or the Communications Society Magazine. Papers dealing with the social economic, political, or legal aspects of the impact of existing, planned, or proposed communication technologies and systems may be sent to Dr. Ralph J. Schwarz, School of Engineering and Applied Science, 510 Mudd Bldg, Columbia University, New York, NY 10027.

The National Tax Association Tax Institute of America announces the 1977 award winners in the annual competition for outstanding doctoral dissertations in government finance and taxation. The $1,000 first prize award was presented to Richard F. Dye, University of Michigan, "Personal Charitable Contributions: Tax Effects and Other Motives." Honorable mention awards of $500 each went to Robert D. Norton, Princeton University, "City Life Cycles and American Urban Policy," and Thomas Schneeweis, University of Iowa, "Municipal Bond Ratings and Market Determined Risk Measures." The members of the 1977 selections committee were Professors George F. Break, Arthur D. Lynn, Jr., Charles E. McLure, Jr., Oliver Oldman, and James A. Papke. Information on the 1978 award competition may be obtained from Professor James A. Papke, Department of Economics, Krannert Graduate School of Management, Purdue University, West Lafayette, Indiana 47907.

Papers are invited for the second conference on major international economic issues to be held at the University of Southern California, Feb. 1-3, 1979. The theme of the conference: How to Eradicate Poverty in the Rich and the Poor Countries. Please send proposals to Professor Nake M. Kamrany, Department of Economics, University of Southern California, Los Angeles, CA 90007.

Deaths

Dorothy S. Brady, professor emeritus, department of economics, University of Pennsylvania, May 17, 1977.

Peter J. Kalman, professor of economics and mathematics, State University of New York-Stony Brook (visiting professor, Harvard University), Cambridge, Mass., Aug. 31, 1977.

James G. Witte, professor of economics, Indiana University, July 17, 1977.

Retirements

Joseph A. Batchelor, department of economics, Indiana University, May 1977.

Tom Bullock Hyder, associate professor of economics, North Texas State University, Aug. 31, 1977.

Visiting Foreign Scholars

John S. Lane, London School of Economics: visiting foreign scholar, department of economics, University of California-San Diego, July 1977-June 1978.

Rowland Maddock, University of Wales: visiting professor, department of economics, Carleton College, Aug. 1977 June 1978.

Muhamed Malallah, University of Jordan: visiting professor of world business, American Graduate School of International Management, June 1, 1977.

Promotions

Eugene Bond: professor of accounting, American Graduate School of International Management, July 1, 1977.

Hui-Shyong Chang: associate professor of economics, University of Tennessee, Sept. 1, 1977.

Terrence M. Clauretie: associate professor of economics, Shepherd College, Aug. 19, 1977.

Ronald G. Ehrenberg: professor of economics and labor economics, Cornell University, Nov. 1, 1977.

Ralph F. Frasca: associate professor of economics, University of Dayton, Aug. 15, 1977.

Constantine Glezakos: professor, department of economics, California State University-Long Beach, Sept. 1977.

Maury L. Harris: chief, Monetary and Finance Division, Federal Reserve Bank of New York, Sept. 15, 1977.

Bryan Heathcotte: associate professor of finance, American Graduate School of International Management, July 1, 1977.

E. William Johnson: associate professor of economics, Shepherd College, Aug. 19, 1977.

Herbert J. Kiesling: professor, department of economics, Indiana University, July 1, 1976.

Mordechai E. Lando: chief, Social Surveys Branch, Office of Research and Statistics, Social Security Administration, May 1977.

John Lindholtz: professor of marketing, American Graduate School of International Management, July 1, 1977.

Fredric C. Menz: associate professor of economics, Clarkson College of Technology, July 1, 1977.

James Mills: associate professor of economics, American Graduate School of International Management, July 1, 1977.

James C. Moore: professor of economics, Purdue University, Aug. 1977.

Donald O. Parsons: professor of economics, Ohio State University.

Edward J. Ray: professor of economics, Ohio State University.

Wallace Reed: associate professor of accounting, American Graduate School of International Management, July 1, 1977.

Byron G. Spencer: professor of economics, McMaster University, July 1977.

Richard M. Thornton: associate professor, department of economics, DePaul University, Sept. 1977.

Administrative Appointments

Raveendra N. Batra: chairman, department of economics, Southern Methodist University, Aug. 1, 1977.

Michael E. Borus, Michigan State University: professor of labor and human resources, and director, Center for Human Resource Research, Ohio State University, Sept. 1977.

W. Richard Bossert: chairman, department of world business, American Graduate School of International Management, July 1, 1977.

Henry N. Goldstein: head, department of economics, University of Oregon, Sept. 15, 1977.

J. T. Scott: coordinator of international programs, College of Agriculture, Iowa State University, July 1, 1977.

Mordechai Shechter: chairman, department of economics, University of Haifa, Oct. 1, 1977.

C. William Suver, St. Martin's College: dean, School of Business, Gonzaga University, Aug. 22, 1977.

William R. Waters: chairman, department of economics, DePaul University, July 1, 1977.

Sidney Wertimer, Jr.: provost, Hamilton College, Aug. 1, 1977.

New Appointments

H. Sonmez Atesoglu, International Monetary Fund: assistant professor of economics, Clarkson College of Technology, Sept. 1, 1977.

John W. Ball: instructor, department of economics, Iowa State University, Sept. 1, 1977.

Kenneth Bernauer: economist, Industrial Economies Division, Federal Reserve Bank of New York, Aug. 31, 1977.

Stanley E. Boyle: visiting professor of economics, University of South Carolina, Aug. 16, 1977.

Ralph J. Brown: professor of economics, University of South Dakota, Aug. 1977.

Roland Buck, Texas A&M University: visiting assistant professor, department of economics, Ohio State University.

Chong-nin J. Chan: assistant professor, department of economics, John Carroll University, Sept. 1977.

Carl Chen: assistant professor of economics and finance, University of Dayton, Aug. 15, 1977.

Raymond Cohn, University of Oregon: assistant professor of economics, Illinois State University, Aug. 1977.

David A. Conn, Miami University (Ohio): visiting assistant professor, department of economics, Ohio State University.

Richard Cornwall: adjunct professor, Middlebury College, Fall 1977.

Patrick S. Cotter: assistant professor, department of economics, John Carroll University, Sept. 1977.

Robert Deaver: economist, Public Information Division, Federal Reserve Bank of New York, Oct. 31, 1977.

Robert A. Driskill, Johns Hopkins University: assistant professor, department of economics, Ohio State University.

Robert D. Foster, Dalhousie University: associate professor of economics, Louisiana Tech University, Sept. 6, 1977.

Marc P. Freiman, Congressional Budget Office: assistant professor, department of economics, Wayne State University, Sept. 1977.

N. Gail Frey, California State University: assistant professor, department of economics, Ohio State University.

Irving Gershenberg: associate professor, department of economics, Boston State College, Sept. 1977.

Reuven Glick: economist, Balance of Payments Division, Federal Reserve Bank of New York, Sept. 28, 1977.

Lawrence H. Hadley: assistant professor of economics, University of Dayton, Aug. 15, 1977.

Terry Halpin, Michigan State University: assistant professor of economics, Illinois State University, Aug. 1977.

Alan J. Harrison: assistant professor, department of economics, McMaster University, Aug. 1977.

Harmon H. Haymes, Virginia Commonwealth University: professor of economics, St. Mary's College of Maryland.

Glenn V. Henderson, Jr., Arizona State University: associate professor of finance, Louisiana Tech University, Sept. 6, 1977.

Janet Conrad Hunt: visiting assistant professor of economics, University of South Carolina, Aug. 16, 1977.

John Hurd II: visiting lecturer, department of economics, Williams College, Sept. 1977.

Keith Ihlanfeldt: assistant professor of economics, University of Dayton, Aug. 15, 1977.

John L. Jurewitz: assistant professor, department of economics, Williams College, Sept. 1977.

Hilda Kahne: professor, department of economics, Wheaton College, July 1, 1977.

M. A. H. Katouzian: visiting associate professor, department of economics, McMaster University, July 1977.

John F. Kennan: visiting assistant professor, department of economics, McMaster University, July 1977.

Christopher Klisz: lecturer, department of economics, Wayne State University, Sept. 1977.

Jesse M. Levy: economist, Social Surveys Branch, Office of Research and Statistics, Social Security Administration, July 1977.

Peter J. McCabe, McMaster University: assistant professor of economics, Purdue University, Aug. 1977.

William McNaught: associate economist, The Rand Corporation, Washington, D.C., Aug. 1977.

Shohreh Majin: lecturer, department of economics, Wayne State University, Sept. 1977.

Lawrence A. Mayer, *Fortune Magazine*: advisor, research and statistics function, Federal Reserve Bank of New York, Jan. 1, 1977.

David Molnar: assistant professor of economics, Simmons College, July 1, 1977.

Josef Molsberger, University of Cologne, Germany: professor of economics, University of Tuebingen, Germany, Aug. 11, 1977.

Craig T. Moore: economist, Industrial Economics Division, Federal Reserve Bank of New York, Sept. 12, 1977.

John K. Mullen: assistant professor of economics, Clarkson College of Technology, Sept. 1, 1977.

Carol D. Norling, Illinois State University: assistant professor, department of economics, Wayne State University, Sept. 1977.

Martha Paas: fellow in economics, Carleton College, Aug. 1977.

Richard A. Palfin, Western Washington University: assistant professor, department of economics, Whitman College, Sept. 1977.

Jan Scott Palmer, Denison University: visiting assistant professor, department of economics, Ohio State University.

Coleen Carey Pantalone, Iowa State University: assistant professor of economics, Clarkson College of Technology, Sept. 1, 1977.

Joel L. Prakken: economist, Business Conditions Division, Federal Reserve Bank of New York, Oct. 3, 1977.

Michael D. Rabin: research assistant, economics department, The Rand Corporation, July 1977.

Rati Ram, University of Chicago: assistant professor of economics, Illinois State University, Aug. 1977.

Barbara Rovello: professional staff, Hudson Institute, June 1977.

Rose M. Rubin, Mississippi State University: assistant professor, department of economics, North Texas State University.

Javier Ruiz-Castillo, State University of New York-Stony Brook: visiting lecturer in economics, Purdue University, Aug. 1977.

Krishan G. Saini: economist, Developing Economies Division, Federal Reserve Bank of New York, June 27, 1977.

Robert R. Schneider: assistant professor, department of economics, Williams College, Sept. 1977.

Bruce A. Seaman, University of Chicago: assistant professor, Georgia State University, Jan. 1978.

Louis Silvia, Michigan State University: lecturer, department of economics, Wayne State University, Sept. 1977.

Dennis E. Smallwood: economist, economics department, The Rand Corporation, July 1977.

Kevin C. Sontheimer, State University of New York-Buffalo: visiting research professor, department of economics, University of Pittsburgh, Sept. 1977-Apr. 1978.

William P. Starnes: assistant professor, department of economics, Williams College, Sept. 1977.

Michael L. Visscher, Carnegie-Mellon University: assistant professor, department of economics, Ohio State University.

Jimmy Wayne Wheeler, Florida International University: professional staff, Hudson Institute, Oct. 1977.

Robert L. Welch, University of California-Santa Barbara: assistant professor, department of economics, Wayne State University, Sept. 1977.

G. Thomas West, University of Virginia: visiting assistant professor of economics, Virginia Commonwealth University.

Roberton C. Williams, Jr.: assistant professor, department of economics, Williams College, Sept. 1977.

George E. Wright: assistant professor, department of economics, DePaul University, Sept. 1977.

Tamara M. Woroby: assistant professor, department of economics, McMaster University, July 1977.

Michael Yokell, University of California-Berkeley: senior economist, Solar Energy Research Institute.

Leaves for Special Appointment

Lascelles Anderson, University of Akron: education economist, Harvard Institute for International Development, Harvard University, 1977 78.

Ernst Baltensperger, Ohio State University: Swiss National Bank, Switzerland.

Charles C. Cox, Ohio State University: research fellow, Hoover Institute, Stanford University.

Elizabeth Erickson, University of Akron: U.S. Department of Agriculture, Sector Analysis Division, Jan. 1977 Sept. 1978.

Marshall I. Goldman, Wellesley College: Fulbright-Hayes visiting professor, Moscow State University, *USSR*, Fall 1977.

Patricia H. Kuwayama: consultant, Bank for International Settlements, Basle, Switzerland, Nov. 1977-July 1978.

Efthimios Pournarakis, University of Akron: head, graduate program, Salonica Graduate School of Industrial Studies, Salonica, Greece, Sept. 1977-June 1978.

Frederick C. Schadrack: deputy director, Division of Banking Supervision and Regulation, Board of Governors of the Federal Reserve System, Aug. 1977-July 1979.

Donald R. Sherk, Simmons College: international economist, Office of International Development Banks, U.S. Treasury, Sept. 1, 1977.

Simón Teitel, Inter-American Development Bank and Catholic University of America: visiting fellow, Economic Growth Center, Yale University, 1977-78.

Resignations

Michael Landsberger, Technion-Israel Institute of Technology: University of Haifa, Oct. 1, 1977.

Arie Melnick, Techninon-Israel Institute of Technology: University of Haifa, Oct. 1, 1978.

Donald W. Ramey, DePaul University, Summer 1977.

Rubin Saposnik, Georgia State University: University of Kentucky, Sept. 1977.

Larry D. Schroeder, Georgia State University: Syracuse University, Sept. 1977.

Inderjit Singh, Ohio State University: World Bank, June 1977.

Abraham Subotnik, Technion-Israel Institute of Technology: University of Haifa, Oct. 1, 1978.

Ronald S. Warren, Jr., U.S. Department of Labor: assistant professor of economics, University of Virginia, Sept. 1, 1977.

John Weicher, Ohio State University: Urban Institute, Sept. 1977.

Daniel Wiseccarver, Ohio State University: Harvard Institute for International Development, Harvard University, June 1977.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, John W.), his present place of employment or enrollment: his new title (if any), his next place of employment (if known or if changed), and the date at which the change will occur.

C. Type each item on a separate 3x5 card, and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, November 1; *June*, February 1; *September*, May 1; *December*, August 1.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the assistant editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

THE AMERICAN ECONOMIC REVIEW

VOL. 68 NO. 2 MAY 1978

PAPERS AND PROCEEDINGS

OF THE

Ninetieth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

New York, New York

December 28–30, 1977

Program Arranged by JACOB MARSCHAK AND JACK HIRSHLEIFER

Papers and Proceedings Edited by GEORGE H. BORTS

AND JAMES A. HANSON

CONTENTS

PAPERS

PROCEEDINGS

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninetieth annual meeting of the American Economic Association. The *Proceedings* consist of the record of the business activities of the Association: the annual membership meeting; the March and December meetings of the Executive Committee; and reports of various Association officers and committees. As with the Notes section in each issue of the *American Economic Review*, they are published to keep the members informed and encourage them to participate in the Association's affairs.

The *Papers* constitute the greater part of this volume. They are roughly equivalent to two regular issues of the *American Economic Review*, but are published under different procedures. About a year in advance, the Association's President-elect (in 1977 Jacob Marschak, in 1978 Robert Solow) acting as program chairman, decides on the topics of sessions at which papers will be presented. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. The program chairman also sets limits on the length of papers at various sessions, and invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the topic of the session, and asks others to give comments on the papers. Some of the sessions are devoted to contributed rather than invited papers. The program chairman decides at the time of organization which sessions are to be printed. The papers to be published are sent to the editorial office of the *Review* about a month before the meetings; the editors of the *Papers and Proceedings* check them for length and content, and send the authors comments and suggestions.

In 1977 there were two significant departures from the most recent issues. First it was the desire of Jacob Marschak, the program chairman, that wherever possible comments be published together with the papers. To accommodate the space required for the comments, the length of the papers was severely restricted to 3500, 3000, or 2500 words, depending on the number of papers in a given session. Not all comments were published. Some arrived too late for the printing deadline, some were delivered orally at the meetings and not written up for publication. In deciding which comments to publish, we relied heavily on recommendations made by the session chairmen. The second departure is that the program chairman decided to publish only invited papers.

The rules under which these papers are published are quite different from those governing articles appearing in regular issues of the *Review*. Their length is strictly controlled. Their content and range of subject matter reflects the wishes of the program chairman to explore and expose the current state of economic research and thinking. In many cases they are exploratory and discursive rather than definitive presentations of research findings. While we do edit the papers to improve content and style, to satisfy space constraints, and to eliminate repetition, we do not subject the papers to any refereeing process, and publication of any paper received prior to the printing deadline that satisfies space requirements is virtually guaranteed.

We would refuse to publish a paper if we concluded after reading it that it was utterly without merit; no paper has yet been rejected on these grounds. The Executive Committee has established another ground for rejection: if a paper cannot be cut to meet space requirements, we may ask the author to authorize its consideration for publication in a regular issue of the *Review* (subject to the usual refereeing process). Or the author may be asked to withdraw the paper and submit it elsewhere.

These policies serve a number of important purposes: The papers can be published without the long delays imposed

by the refereeing process. They are short papers, covering a wide variety of subjects, and in most cases can be understood by nonspecialists. Authors receive a chance to report on research just completed, discuss topical subjects in an informal way, and to summarize longer forthcoming publications. Readers get a chance to browse among a large number of articles which are outside their major areas of interest, but which are not as specialized or as technical as those sometimes found in the regular journals. And while the papers are not refereed, they do provide an accurate picture of the state of thinking in many of the fields of economics.

As many are aware, Jacob Marschak, the program chairman and President-elect for 1977, died suddenly on July 27, 1977. He was responsible for organizing the program of these meetings. During the year Marschak received considerable assistance from his colleague Jack Hirshleifer. After Marschak's death Hirshleifer served as program chairman. In his introductory note to the program Hirshleifer characterized it as follows:

> "Professor Marschak chose three main themes for the 1977 American Economic Association meetings: (1) the boundaries of economics, with several sessions on how economics has related and might relate to other scientific and humanistic disciplines; (2) macroeconomics, with emphasis both upon the goals and the effectiveness of policy; and (3) the application of economics to societal problems. The proceedings will constitute for his many friends, students, and colleagues, a memorial of the achievements and influence of Jacob Marschak."

GEORGE H. BORTS
JAMES A. HANSON

Jacob Marschak, 1898–1977

Jacob Marschak's great contributions to economics and to social science are of two kinds: First, through his keen perception, his original thought, his research, writing, and lectures. Second, through being a natural leader of others, and both an initiator and a catalyst of research discussions. Students, staff members, and colleagues were drawn to him by his intellectual honesty, his ability to involve those around him, his wit, and his sharp insights into ideas and men.

The two contributions are two closely connected aspects of one whole man. Even so, with changing circumstances, the relative emphasis on his own thought and on inspiring others has fluctuated.

The Chicago period in Marschak's life was one that brought out his leadership qualities in full strength. In my account of that period of Marschak's scientific life I will not try to separate the two aspects of his role. I will actually start a little earlier, from soon after Marschak joined the faculty of the New School for Social Research in 1939.

In those difficult years, Marschak organized, and was the soul of, a seminar on econometric methods and results that met periodically, drawing its regular participants from the greater New York area, with occasional visitors from more distant centers. The participants and visitors included Sidney Alexander, Franz Alt, Trygve Haavelmo, Carl Kaysen, myself, Wassily Leontief, Franco Modigliani, Albert Neisser, Paul Samuelson, Joseph Schumpeter, Abraham Wald, and probably several others whose names I have been unable to recall or trace.

In that period explicit formulation of a full set of quantitative relations of economic behavior was still met with reserve if not suspicion. To its participants, therefore, the seminar was an oasis of econometric discussion in the New York area. In the words of one of them, Marschak instilled in the seminar a strong penchant for the fundamental problems. The seminar met on the premises of the National Bureau of Economic Research, at Columbus Circle and at Hillside, thanks to the hospitality of its Director and staff.

The contacts between Marschak, Haavelmo, and Wald were particularly intensive in this period. Their interactions came to fruition in three fundamental papers that appeared within the span of one and one-half years, the first one by Haavelmo, "On the Statistical Implications of a System of Simultaneous Equations" (January 1943), the next by Mann and Wald, "On the Statistical Treatment of Stochastic Difference Equations," (July–October, 1943), and the third by Marschak and Andrews on "Random Simultaneous Equations and the Theory of Production," (July–October 1944.)

These papers knitted together three distinct fields of inquiry. Between them, the authors took from economic theory the explanation of the determination of economic quantities, prices, or values by a system of simultaneous equations. They rebuilt and generalized the methods of statistical regression to take the joint determination of economic variables into account. Finally, the repetitiousness over time of these interactions of prices and quantities made it possible to apply the statistical theory of stochastic difference equations.

To my knowledge the Marshak-Andrews paper was the first in which the new approach proposed by Haavelmo was applied to estimate economic behavior relationships. It was followed in 1947 by a paper by Haavelmo, and another jointly by Haavelmo and Girshick, applying the new methods to estimate the marginal propensity to consume in the former paper, and demand and supply functions for food in the latter.

Meanwhile, the scene had shifted to the University of Chicago, to which Marschak moved in 1943 and where he assumed the Directorship of the Cowles Commission for

Research in Economics. In that role Marschak's vision and persistence led to a systematic development and application of the new methods to the econometrics of markets and of macroeconomic stability or fluctuations. By a process mixing empathy with shrewd selection, he assembled an unavoidably somewhat rotating cross-disciplinary staff, from among whom I mention Theodore Anderson, Kenneth Arrow, Herman Chernoff, Carl Christ, Evsey Domar, Trygve Haavelmo, Leonid Hurwicz, Lawrence Klein, myself, Don Patinkin, Herman Rubin, and Sam Schurr.

Marschak himself contributed characteristically pithy and terse opening chapters to the two methodological volumes that were a part of this development. These chapters supplied a framework and direction for the interrelated efforts of the members of the team. The principal application carried through in that period was Lawrence Klein's volume "Economic Fluctuations in the United States, 1921–41," which became an important bridge between Tinbergen's pioneering models from just before the war and the extended development of econometric modeling of the present day—a symbiosis of academic research projects and a modeling industry that appears to be meeting the market test.

As always, so in the Chicago period Marschak attracted members of neighboring institutions who took part in the research discussions. Abba Lerner, then of Roosevelt College, Franco Modigliani, while at the University of Illinois, and Herbert Simon, then at the Illinois Institute of Technology, were frequent participants.

As time went on, the interests of the group diversified, and Marschak intensified his concerns with the theory of money and the theory of decisions under uncertainty by individuals and organizations that have occupied him throughout his career. His fundamental contributions in these areas will be described by Kenneth Arrow.

But I still want to share with you my personal experiences of Jascha Marschak, the man himself, as revealed in conversations and in actions.

He was a citizen of the World, whose career extended over four countries. The earliest activity that he described to me was his service, as a very young man, as Minister of Labor in the short-lived Coalition Government of the Terek Republic in the North Caucasus (predominantly a Menshevik, in European terms a Social-Democratic, regime).

Jascha, conveying a sense of the tragicomic aspects of that experience, described that government to me as a paedocracy (paidokratia, a government of children). He also related a warning of its impending overthrow that he had received from a person with a longer local experience. This man had warned: "When the corn has grown high enough to conceal a man on horseback, the government will fall." That, Jascha said, is what happened.

Most of Jascha's professional contributions were made in the Weimar Republic of Germany, in England in the years before the World War II, and in the United States during and after the war. In a conversation about the experience of changing countries of residence, he told me that, in each case, after living two or three years in a new country he had started to identify with its people and its modes of life.

Just the same he remained very Russian and very Jewish throughout these changes. He felt strong attachments to family, relatives, and friends. He spoke to me of what he had learned from long conversations with his older brother Leon, an engineer and scientist. He also spoke of his friendship and admiration for Wladimir Woytinski, a somewhat older Russian economist, whose life and work Jascha has described in a short but loving paper reprinted in Volume III of his *Selected Essays*.

Jascha had a gentle and humane wit, often in the form of stories about mostly unnamed people's particularities or foibles—stories that always made a point. Sometimes his humor was at his own expense. At the meeting of the Econometric Society in Namur, Belgium, held in 1935, a participant criticized some aspects of his book "Die Elastizität der Nachfrage" (The Elasticity

of Demand). In response Jascha said that he was no longer satisfied with that book. In fact, he said (no doubt exaggerating), he would have been relieved if the book had been included in the book burnings by the Nazis. A German participant got up and left the room.

Another aspect of Jascha's wit was his delight in logical twists. This came out in his story of a quarrel between two groups of pupils associated with two leading Rabbis in medieval Prague. It is a story I have cherished and retold. A pupil of one group said: "Our Rabbi is a holy man, he speaks with God." One of the other group retorted: "Your Rabbi is a liar." From the first group came the response: "Would God speak to a liar?"

Throughout his varied professional life, Jascha was strongly and personally concerned with his students, with his collaborators and with his colleagues. He would encourage the timid but turn away from the pretentious. In particular, in discussions of quantitative problems, he was wary of participants who were reluctant to use the blackboard. He had no patience with careerism, or even with just making bows of respect to prevailing doctrines or opinions.

To me, personally, he has been a shining example and a patient mentor. I have been very fortunate in having known him well.

TJALLING C. KOOPMANS
Yale University
December 29, 1977

Jacob Marschak's Contributions to the Economics of Decision and Information

Scholars, like all other individuals, vary greatly in their tolerance for uncertainty and ambiguity. Some feel no comfort until they have a theoretical framework or at least a vision capable of explaining to their satisfaction the phenomena of interest in their field. The entertainment of alternative hypotheses is difficult for them. Others can contemplate with equanimity the possibility that our empirical knowledge and theoretical understanding are compatible with more than one view of the world, that only gradually will there be greater resolution.

Both types of scholars have their roles, and it is just as well that both types exist. The demand for certainty is a powerful incentive to developing the theory and empirical investigations that push the subject forward. The risk-tolerant scholar is more open to new ideas and, in particular, is liable to play a special role in synthesis, in the yoking together of ideas from disparate fields.

Jacob Marschak certainly belongs on the risk-tolerant end of the spectrum. No one who knew him failed to be struck by the intensity of his desire for new knowledge. I remember my first meeting with Jacob Marschak at a seminar held, as I recall, at the New School for Social Research in the winter 1941–42. In the course of discussion I made some rash remarks about the unimportance of multicollinearity as a statistical problem. Several months later we met again at a picnic. While munching on hot dogs he insisted on cross-examining my position with great pertinacity. The idea that a senior, distinguished economist was seeking knowledge from a very young graduate student was quite an overwhelming experience.

This pattern was repeated in the free-for-all staff meetings of the Cowles Commission. The participants took the conduct of the meetings as normal and hardly noticed its special flavor, but when Norman Kaplan, a graduate student not part of the Cowles Commission circle, came to a staff meeting one day, he was astounded and perhaps even appalled at the freedom with which the youngest graduate student corrected Jascha.

It is perhaps then not entirely surprising that Marschak's sense of the uncertainties with which the world abounds should translate itself into a scholarly interest in decision making under uncertainty and the problems of communication and information. As befits an information theorist, Marschak's scientific behavior depended not only on the shape of his utility function but also on observations on the external world. The power of the socialist movement and ideology in Europe, reinforced by the Russian revolution, in which Marschak had personal experiences, gave rise to a great debate on the viability of socialism as an economic system. The quality of the debate was of the highest order and attracted some of the finest minds in economics: Pareto, Barone, von Mises, Schumpeter, von Hayek, Lange, and Lerner. It is sometimes forgotten that Marschak contributed a significant paper to this discussion, emphasizing the value of a competitive price system in achieving an efficient allocation of resources and arguing that it was more likely to be achieved under socialism than under monopolistic capitalism.

Thus, the theme of economic organization was already present in Marschak's mind in an early period. In the 1930's, Marschak, following the work of Keynes, Hicks, and others became interested in the demand for money as a crucial part of the economic system. Following earlier hints he sought to ground this demand in the theory of behavior under uncertainty. Though he went further in this regard than

anyone else, the rejection of cardinal utility theory, at that time universal among the *avant-garde*, made the construction of the theory difficult.

With the introduction of new concepts and analytical tools, by von Neumann, Morgenstern, and Wald, it was possible to discuss these matters more precisely and Marschak returned to the analysis for the demand for money with his 1949 and 1950 papers. However, his interests had broadened by this time. First of all, he developed important expositions of the new theories of expected utility and of subjective probability, as developed by Ramsey and by Savage. At the same time Marschak was alert to the contemporaneous developments of learning theory and psychology. In several important papers he helped the synthesis of the two approaches. He even pioneered in the use of experimental methods in these economic situations.

The greater novelty however was the synthesis of statistical decision theory, the economics of decision making under uncertainty and the theory of organization, a combination to which he gave the name of the *theory of teams*. He had a grand vision which represents a new and sophisticated interpretation of the problem of economic coordination, so central to the debate on socialism and indeed in the development of economic thought from the days of Adam Smith on. He recognized fully Hayek's emphasis on the dispersed nature of knowledge in the economy.

Imagine then an economic system composed of many agents. Each agent makes decisions, and the outcome for the economy as a whole depends on the decisions made and on certain facts about the world. These facts might be interpreted as the realization of a random variable of many dimensions. Each individual initially has a limited knowledge about the world. Indeed the economy as a whole may have limited knowledge, but the important point is that each agent has some information which other agents do not have.

Assume away, at least for the time being, the fact that the agents probably have different interests. Assume they have the same goals. They must then jointly adopt decision rules, so that each individual agent makes a decision on the basis of the information available to him. This information is a random variable at the time the decision rules are drawn up.

One can add to this system the possibility of communication. That is, individuals may transmit some of the information they have to other individuals. It is of the essence of course that communication is scarce and the formulation of the team problem must in some way or another impose a cost or a limitation on the amount of communication. This reflects the basic reason why knowledge in a society is so dispersed.

This general point of view was set forth in two important papers in 1954 and 1955 and has been richly developed since by Marschak in collaboration with Radner and also by Radner, Groves, Jordan and others. This general formulation of the problem gives rise to several subsidiary problems of which the most important is a clearer grasp of the economics of information. Information is not quantitatively measured and differs therefore from ordinary scarce commodities, such as sugar. The communications engineers, most notably Shannon, have developed their measures but their economic appropriateness is not always clear. Marschak's papers have explored in detail the relations between the entropy-like measures and those more definitely based on economic criteria of cost scarcity and benefit. The partial ordering of economic states through the concept of efficiency has as its analogue a partial ordering of information structures, and Marschak's work with Miyasawa showed how this can be done, drawing on the earlier work of Blackwell.

Marschak's synthetic aims brought him into contact with a wide variety of disciplines: not only economists and statisticians, but also communications engineers, logicians, psychologists, management scientists, and even physicists and

biologists. The impulse that he gave to thinking in the area of information and communication came not only from his own work but from his personal leadership as especially exemplified in the interdisciplinary seminar at the University of California at Los Angeles.

Like few others, his life and work were centrally located in the social matrix of inquiry. His rare intellectual spirit and rich personality will remain with us.

KENNETH ARROW
Harvard University
December 29, 1977

Rationality as Process and as Product of Thought

By HERBERT A. SIMON*

This opportunity to deliver the Richard T. Ely Lecture affords me some very personal satisfactions. Ely, unbeknownst to him, bore a great responsibility for my economic education, and even for my choice of profession. The example of my uncle, Harold Merkel, who was a student of Commons and Ely at Wisconsin before World War I, taught me that human behavior was a fit subject for scientific study, and directed me to economics and political science instead of high energy physics or molecular biology. Some would refer to this as satisficing, for I had never heard of high energy physics or molecular biology, and hence was spared an agonizing weighing of alternative utiles. I simply picked the first profession that sounded fascinating.

Ely's influence went much further than that. My older brother's copy of his *Outlines of Economics*—the 1930 edition—was on our bookshelves when I prepared for high school debates on tariffs versus free trade, and on the Single Tax of Henry George. It provided me with a sufficiently good grounding in principles that I was later able to take Henry Simons' intermediate theory course at the University of Chicago, and the graduate courses of Frank Knight and Henry Schultz without additional preparation.

The Ely textbook, in its generation, held the place of Samuelson or Bach in ours. If it would not sound as though I were denying any progress in economics over the past half century, I might suggest that Ely's textbook could be substituted for any of our current ones at a substantial reduction in weight, and without students or teacher being more than dimly aware of the replacement. Of course they would not hear from Ely about marginal propensities to do this or that, nor about the late lamented Phillips curve. But monetarists could rejoice in Ely's uncompromising statement of the quantity theory (p. 298, italics), and in his assertion that "the solution of the problem of unemployment depends largely upon indirect measures, such as monetary and banking reform"—Ely does go on to say, however, that "we shall recognize that society must offer a willing and able man an opportunity to work" (p. 528).

I. Rationality in and out of Economics

I have more than personal reasons for directing your attention to Ely's textbook. On page 4, we find a definition of economics that is, I think, wholly characteristic of books contemporary with his. "Economics," he says, "is the science which treats of those social phenomena that are due to the wealth-getting and wealth-using activities of man." Economics, that is to say, concerns itself with a particular subset of man's behaviors—those having to do with the production, exchange, and consumption of goods and services.

Many, perhaps most, economists today would regard that view as too limiting. They would prefer the definition proposed in the *International Encyclopedia of the Social Sciences*: "Economics . . . is the study of the allocation of scarce resources among unlimited and competing uses" (vol. 4, p. 472). If beefsteak is scarce, they would say, so are votes, and the tools of economic analysis can be used as readily to analyze the allocation of the one as of the other. This point of view has launched economics into many excursions and incursions into political science and her other sister social sciences, and has generated a certain amount of hubris in the profession with respect to its broader civilizing mission. I

*Carnegie-Mellon University.

would suppose that the program of this meeting, with its emphasis upon the relations between economics and the other social sciences, is at least partly a reflection of that hubris.

A. *Rationality in Economics*

The topic of allocating scarce resources can be approached from either its normative or its positive side. Fundamental to the approach from either side are assumptions about the adaptation of means to ends, of actions to goals and situations. Economics, whether normative or positive, has not simply been the study of the allocation of scarce resources, it has been the study of the *rational* allocation of scarce resources.

Moreover, the term "rational" has long had in economics a much more specific meaning than its general dictionary signification of "agreeable to reason; not absurd, preposterous, extravagant, foolish, fanciful, or the like; intelligent, sensible." As is well known, the rational man of economics is a maximizer, who will settle for nothing less than the best. Even his expectations, we have learned in the past few years, are rational (see John Muth, 1961).[1] And his rationality extends as far as the bedroom for, as Gary Becker tells us, "he would read in bed at night only if the value of reading exceeded the value (to him) of the loss in sleep suffered by his wife" (1974, p. 1078).

It is this concept of rationality that is economics' main export commodity in its trade with the other social sciences. It is no novelty in those sciences to propose that people behave rationally—if that term is taken in its broader dictionary sense. Assumptions of rationality are essential components of virtually all the sociological, psychological, political, and anthropological theories with which I am familiar. What economics has to export, then, is not rationality, but a very particular and special form of it—the rationality of the utility maximizer, and a pretty smart one at that. But international flows have to be balanced. If the program of this meeting aims at more active intercourse between economics and her sister social sciences, then we must ask not only what economics will export, but also what she will receive in payment. An economist might well be tempted to murmur the lines of the tentmaker: "I wonder often what the Vintners buy—One half as precious as the stuff they sell."

My paper will be much concerned with that question, and before I proceed, it may be well to sketch in outline the path I propose to follow in answering it. The argument has three major steps.

First, I would like to expand on the theme that almost all human behavior has a large rational component, but only in terms of the broader everyday sense of rationality, not the economists' more specialized sense of maximization.

Second, I should like to show that economics itself has not by any means limited itself to the narrower definition of rationality. Much economic literature (for example, the literature of comparative institutional analysis) uses weaker definitions of rationality extensively; and that literature would not be greatly, if at all, improved by substituting the stronger definition for the weaker one.[2] To the extent that the weaker definition is adequate for purposes of analysis, economics will find that there is indeed much that is importable from the other social sciences.

Third, economics has largely been preoccupied with the *results* of rational choice rather than the *process* of choice. Yet as economic analysis acquires a broader concern with the dynamics of choice under uncertainty, it will become more and more essential to consider choice processes. In the past twenty years, there have been im-

[1]The term is ill-chosen, for rational expectations in the sense of Muth are profit-maximizing expectations only under very special circumstances (see below). Perhaps we would mislead ourselves and others less if we called them by the less alluring phrase, "consistent expectations."

[2]For an interesting argument in support of this proposition from a surprising source, see Becker (1962). What Becker calls "irrationality" in his article would be called "bounded rationality" here.

portant advances in our understanding of procedural rationality, particularly as a result of research in artificial intelligence and cognitive psychology. The importation of these theories of the processes of choice into economics could provide immense help in deepening our understanding of the dynamics of rationality, and of the influences upon choice of the institutional structure within which it takes place.

We begin, then, by looking at the broader concept of rationality to which I have referred, and its social science applications.

B. *Rationality in the Other Social Sciences: Functional Analysis*

Let me provide some examples how rationality typically enters into social science theories. Consider first so-called "social exchange" theories (see, for example, George Homans). The central idea here is that when two or more people interact, each expects to get something from the interaction that is valuable to him, and is thereby motivated to give something up that is valuable to the others. Social exchange, in the form of the "inducements-contributions balance" of Chester I. Barnard and the author (1947), has played an important role in organization theory, and in even earlier times (see, for example, George Simmel) was a central ingredient in sociological theories. Much of the theorizing and empirical work on the topic has been concerned with determining what constitutes a significant inducement or contribution in particular classes of exchange situations—that is, with the actual shape and substance of the "utility function." Clearly, the man of social exchange theory is a rational man, even if he is never asked to equate things at the margin.

It is perhaps more surprising to discover how pervasive assumptions of rationality are in psychoanalytic theory—confirming the suspicion that there is indeed method in madness. In his *Five Lectures* Sigmund Freud has this to say about neurotic illnesses:

> We see that human beings fall ill when, as a result of external obstacles or of an internal lack of adaptation, the satisfaction of their erotic needs *in reality* is frustrated. We see that they then take flight into *illness* in order that by its help they may find a satisfaction to take the place of what has been frustrated . . . We suspect that our patients' resistance to recovery is no simple one, but compounded of several motives. Not only does the patient's ego rebel against giving up the repressions by means of which it has risen above its original disposition, but the sexual instincts are unwilling to renounce their substitutive satisfaction so long as it is uncertain whether reality will offer them anything better.

Almost all explanations of pathological behavior in the psychoanalytic literature take this form: they explain the patient's illness in terms of the functions it performs for him.

The quotation from Freud is illustrative of a kind of functional reasoning that goes far beyond psychoanalysis and is widely used throughout the social sciences, and especially anthropology and sociology. Behaviors are functional if they contribute to certain goals, where these goals may be the pleasure or satisfaction of an individual or the guarantee of food or shelter for the members of a society. Functional analysis in this sense is concerned with explaining how "major social patterns operate to maintain the integration or adaptation of the larger system" (see Frank Cancian). Institutions are functional if reasonable men might create and maintain them in order to meet social needs or achieve social goals.

It is not necessary or implied that the adaptation of institutions or behavior patterns to goals be conscious or intended. When awareness and intention are present, the function is usually called *manifest*, otherwise it is a *latent* function. The function, whether it be manifest or latent, provides the grounds for the reasonableness or rationality of the institution or behavior pattern. As in economics, evolutionary arguments are often adduced to explain the persistence and survival of

functional patterns, and to avoid assumptions of deliberate calculation in explaining them.

In practice, it is very rarely that the existence or character of institutions are *deduced* from the functions that must be performed for system survival. In almost all cases it is the other way round; it is empirical observation of the behavior pattern that raises the question of why it persists—what function it performs. Perhaps, in an appropriate axiomatic formulation, it would be possible to *deduce* that every society must have food-gathering institutions. In point of fact, such institutions can be *observed* in every society, and their existence is then rationalized by the argument that obtaining food is a functional requisite for all societies. This kind of argument may demonstrate the sufficiency of a particular pattern for performing an essential function, but cannot demonstrate its necessity—cannot show that there may not be alternative, functionally equivalent, behavior patterns that would satisfy the same need.

The point may be stated more formally. Functional arguments are arguments about the movements of systems toward stable self-maintaining equilibria. But without further specification, there is no reason to suppose that the attained equilibria that are reached will be global maxima or minima of some function rather than local, relative maxima or minima. In fact, we know that the conditions that every local maximum of a system be a global maximum are very strong (usually some kind of "convexity" conditions).

Further, when the system is complex and its environment continually changing (that is, in the conditions under which biological and social evolution actually take place), there is no assurance that the system's momentary position will lie anywhere near a point of equilibrium, whether local or global. Hence, all that can be concluded from a functional argument is that certain characteristics (the satisfaction of certain functional requirements in a particular way) are consistent with the survival and further development of the system, not that these same requirements could not be satisfied in some other way. Thus, for example, societies can satisfy their functional needs for food by hunting or fishing activities, by agriculture, or by predatory exploitation of other societies.

C. *Functional Analysis in Economics*

Functional analysis of exactly this kind, though with a different vocabulary, is commonly employed by economists, especially when they seek to use economic tools to "explain" institutions and behaviors that lie outside the traditional domains of production and distribution. Moreover, it occurs within those domains. As an example, the fact is observed that individuals frequently insure against certain kinds of contingencies. Attitudes are then postulated (for example, risk aversion) for which buying insurance is a functional and reasonable action. If some people are observed to insure, and others not, then this difference in behavior can be explained by a difference between them in risk aversion.

To take a second example, George Stigler and Becker wish to explain the fact (if it is a fact—their empiricism is very casual) that as people hear more music, they want to hear still more. They invent a commodity, "music appreciation" (not to be confused with time spent in listening to music), and suggest that listening to music might produce not only immediate enjoyment but also an investment in *capacity* for appreciating music (i.e., in amount of enjoyment produced per listening hour). Once these assumptions are granted, various conclusions can be drawn about the demand for music appreciation. However, only weak conclusions follow about listening time unless additional strong postulates are introduced about the elasticity of demand for appreciation.

A rough "sociological" translation of the Stigler-Becker argument would be that listening to music is functional both in producing pleasure and in enhancing the pleasure of subsequent listening—a typical functional argument. It is quite unclear what is gained by dressing it in the garb of

marginalism. We might be willing to grant that people would be inclined to invest more in musical appreciation early in life than later in life (because they would have a longer time in which to amortize the investment) without insisting that costs and returns were being equated at the margin, and without gaining any new insights into the situation from making the latter assumption.

A sense of fairness compels me to take a third example from my own work. In my 1951 paper, I defined the characteristics of an employment contract that distinguish it from an ordinary sales contract, and then showed why reasonable men might prefer the former to the latter as the basis for establishing an employment relation. My argument requires a theorem and fifteen numbered equations, and assumes that both employer and employee maximize their utilities. Actually, the underlying functional argument is very simple. An employee who didn't care very much which of several alternative tasks he performed would not require a large inducement to accept the authority of an employer—that is, to permit the employer to make the choice among them. The employer in turn would be willing to provide the necessary inducement in order to acquire the right to postpone his decisions about the employee's agenda, and in this way to postpone some of his decisions whose outcomes are contingent on future uncertain events.[3] The rigorous economic argument, involving the idea of maximizing behavior by employer and employee, is readily translatable into a simple qualitative argument that an employment contract may be a functional ("reasonable") way of dealing with certain kinds of uncertainty. The argument then explains why employment relations are so widely used in our society.

The translation of these examples of economic reasoning into the language of functional analysis could be paralleled by examples of translation scholarship which run in the opposite direction. Political scientists, for example, long ago observed that under certain circumstances institutions of representative democracy spawned a multiplicity of political parties, while under other circumstances, the votes were divided in equilibrium between two major parties. These contrasting equilibria could readily be shown by functional arguments to result from rational voting decisions under different rules of the electoral game, as was observed by Maurice Duverger, in his classic work on political parties, as well as by a number of political scientists who preceded him. In recent years, these same results have been rederived more rigorously by economists and game theorists, employing much stronger assumptions of utility maximization by the voters; it was hard to see that the maximization assumptions have produced any new predictions of behavior.[4]

D. *Summary*

Perhaps these examples suffice to show that there is no such gap as is commonly supposed between the view of man espoused by economics and the view found in the other social sciences. The view of man as rational is not peculiar to economics, but is endemic, and even ubiquitous, throughout the social sciences. Economics tends to emphasize a particular

[3]Recently, Oliver Williamson has pointed out that I would have to introduce slightly stronger assumptions to justify the employment contract as rational if one of the alternatives to it were what he calls a "contingent claims" contract, but the point of my example is not affected. To exclude the contingent claims contract as a viable alternative, we need merely take account of the large transaction costs it would entail under real world conditions.

[4]For an introduction to this literature, see William H. Riker and Peter C. Ordeshook, and Riker. Anthony Downs' book belongs to an intermediate genre. While it employs the language of economics, it limits itself to verbal, nonrigorous reasoning which certainly does not make any essential use of maximizing assumptions (as contrasted with rationality assumptions in the broader sense), and which largely translates into the economic vocabulary generalizations that were already part of the science and folklore of politics. In the next section, other examples of this kind of informal use of rationality principles are examined to analyze institutions and their behavior.

form of rationality—maximizing behavior—as its preferred engine of explanation, but the differences are often differences in vocabulary more than in substance. We shall see in a moment that in much economic discussion the notion of maximization is used in a loose sense that is very close to the common sense notions of rationality used elsewhere in the social sciences.

One conclusion we may draw is that economists might well exercise a certain amount of circumspection in their endeavors to export economic analysis to the other social sciences. They may discover that they are sometimes offering commodities that are already in generous supply, and which can therefore be disposed of only at a ruinously low price. On the other side of the trade, they may find that there is more of interest in the modes and results of inquiry of their fellow social scientists than they have generally been aware.

II. On Applying the Principle of Rationality

What is characteristic of the examples of functional analysis cited in the last section, whether they be drawn from economics or from the other social sciences, is that they are not focused on, or even much concerned with, how variables are equated at the margin, or how equilibrium is altered by marginal shifts in conditions (for example, shifts in a supply or demand schedule). Rather, they are focused on qualitative and structural questions, typically, on the choice among a small number of discrete institutional alternatives:

Not "how much flood insurance will a man buy?" but "what are the structural conditions that make buying insurance rational or attractive?"

Not "at what levels will wages be fixed?" but "when will work be performed under an employment contract rather than a sales contract?"

If we want a natural science analogy to this kind of theorizing, we can find it in geology. A geologist notices deep scratches in rock; he notices that certain hills of gravel are elongated along a north-south axis, and that the boulders embedded in them are not as smooth as those usually found on beaches. To explain these facts, he evokes a structural, and not at all quantitative, hypothesis: that these phenomena were produced by the process of glaciation.

In the first instance, he does not try to explain the depth of the glacial till, or estimate the weight of the ice that produced it, but simply to identify the basic causative process. He wants to explain the role of glaciation, of erosion, of vulcanization, of sedimentation in producing the land forms that he observes. His explanations, morever, are after-the-fact, and not predictive.

A. *Toward Qualitative Analysis*

As economics expands beyond its central core of price theory, and its central concern with quantities of commodities and money, we observe in it this same shift from a highly quantitative analysis, in which equilibration at the margin plays a central role, to a much more qualitative institutional analysis, in which discrete structural alternatives are compared.

In these analyses aimed at explaining institutional structure, maximizing assumptions play a much less significant role than they typically do in the analysis of market equilibria. The rational man who sometimes prefers an employment contract to a sales contract need not be a maximizer. Even a satisficer will exhibit such a preference whenever the difference in rewards between the two arrangements is sufficiently large and evident.

For this same reason, such analyses can often be carried out without elaborate mathematical apparatus or marginal calculation. In general, much cruder and simpler arguments will suffice to demonstrate an inequality between two quantities than are required to show the conditions under which these quantities are equated at the margin. Thus, in the recent works of Janos Kornai, Williamson, and John Montias on economic organization, we find only rather modest and simple ap-

plications of mathematical analysis. In the ways in which they involve principles of rationality, the arguments of these authors resemble James March and the author's *Organizations* more closely than Paul Samuelson's *Foundations*.[5]

What is the predominant form of reasoning that we encounter in these theoretical treatments of social institutions? Do they contain arguments based on maximizing assumptions? Basically, they rest upon a very simple form of causal analysis. Particular institutional structures or practices are seen to entail certain undesirable (for example, costly) or desirable (for example, value-producing) consequences. *Ceteris paribus*, situations and practices will be preferred when important favorable consequences are associated with them, and avoided when important unfavorable consequences are associated with them. A shift in the balance of consequences, or in awareness of them, may motivate a change in institutional arrangements.

Consider the following argument from Montias typical of this genre of analysis, which relates to the balance in organizations between centralization and decentralization.

> Decentralizing measures are generally aimed at remedying two shortcomings of an 'overcentralized' system structure. (1) Superordinates are overburdened with responsibility for the detailed direction and coordination of their subordinates' activities. (2) This 'petty tutelage' deprives subordinates of the opportunity to make decisions that might increase the payoff of the organization of which they are a part. . . . Why not loosen controls . . . ? . . . When controls are loosened, unless the incentive system is modified to bring about greater harmony between the goals of supervisors and supervisees, it may induce producers to shift their input and output mix in directions that . . . vitiate any benefits that might be reaped by the organization as a whole from the exercise of greater initiative at lower tiers. [p. 215]

Here two costs or disadvantages of centralization (burden on supervisors, restriction of choice-set of subordinates) are set off against a disadvantage of decentralization (goals of subordinates divergent from organization goals).

What can we learn about organization from an argument like this? Certainly little or nothing about the optimal balance point between centralization and decentralization in any particular organization. Rather, we might derive conclusions of these kinds:

1. That increasing awareness of one of the predicted consequences may cause an organization to move in the direction of centralization or decentralization. (For example, an egregious case of "suboptimizing" by a subordinate may cause additional centralized controls to be instituted.)

2. That new technical devices may tilt the balance between centralization and decentralization. For example, invention and adoption of divisionalized profit and loss statements led toward decentralization of many large American business firms in the 1950's; while reduction in information costs through computerization led at a later date to centralization of inventory control decisions in those same firms.

Of course Montias' conclusions could also be derived from a more formal optimization analysis—in fact he presents

[5]A notable exception to this generalization about the economic literature on organizations is the work of Jacob Marschak and Roy Radner on the theory of teams. These authors chose the strategy of detailed, precise analysis of the implications of maximizing assumptions for the transmission of information in organizations. The price they paid for this rigor was to find themselves limited to the highly simplified situations where solutions could be found for the mathematical problems they posed. We need not, of course, make an either-or choice between these two modes of inquiry. While it may be difficult or impossible to extend the formal analysis of the theory of teams to problems of real world complexity, the rigorous microtheory may illuminate the workings of important component mechanisms in the complex macrosituations. The methodological issues in choosing between analytic tractability and realism are quite parallel to those involved in the choice between laboratory and field methods for gathering empirical information about social phenomena. Neither one by itself marks the exclusive path toward truth.

such an analysis on the two pages following the passage quoted above. But it is not clear that anything new is added by the formalization, since the parameters imputed to the system are largely unmeasured and unmeasurable.

There is something to be said for an Ockham's Razor that, eschewing assumptions of optimization, provides an explanation of behavior that is consistent with *either* optimizing or satisficing procedures on the part of the human agents. Parsimony recommends that we prefer the postulate that men are reasonable to the postulate that they are supremely rational when either one of the two assumptions will do our work of inference as well as the other.[6]

B. *Procedural Rationality*

The kind of qualitative analysis I have been describing has another virtue. In complex situations there is likely to be a considerable gap between the real environment of a decision (the world as God or some other omniscient observer sees it) and the environment as the actors perceive it. The analysis can then address itself either to normative questions—the whole range of consequences that *should* enter into decisions in such situations—or descriptive questions, including the questions of which components of the situation are likely to be taken into account by the actors, and how the actors are likely to represent the situation as a whole.

In the precomputer era, for example, it was very difficult for managers in business organizations to pay attention to all the major variables affected by their decisions. Company treasurers frequently made decisions about working capital with little or no attention to their impact on inventory levels, while production and marketing executives made decisions about inventory without taking into account impacts on liquidity. The introduction of computers changed the ways in which executives were able to reach decisions; they could now view them in terms of a much wider set of interrelated consequences than before. The perception of the environment of a decision is a function of—among other things—the information sources and computational capabilities of the executives who make it.

Learning phenomena are also readily handled within this framework. A number of the changes introduced into planning and control procedures in eastern European countries during the 1960's were instituted when the governments in question learned by experience of some of the dysfunctional consequences of trying to control production by means of crude aggregates of physical quantities. An initial distrust of prices and market mechanisms was gradually and partially overcome after direct experience of the disadvantages of some of the alternative mechanisms. These learning experiences could be paralleled with experiences of American steel companies, for example, that experimented with tonnage incentives for mill department superintendents.

A general proposition that might be asserted about organizations is that the number of considerations that are potentially relevant to the effectiveness of an organization design is so large that only a few of the more salient of these lie within the circle of awareness at any given time, that the membership of this subset changes continually as new situations (produced by external or internal events) arise, and that "learning" in the form of reaction to perceived consequences is the dominant way in which rationality exhibits itself.

In a world where these kinds of adjustments are prominent, a theory of rational behavior must be quite as much concerned with the characteristics of the rational actors—the means they use to cope with uncertainty and cognitive complexity—as

[6]Ockham is usually invoked on behalf of the parsimony of optimizing assumptions, and against the additional *ad hoc* postulates that satisficing models are thought to require in order to guarantee uniqueness of solutions. But that argument only applies when we are trying to deduce unique equilibria, a task quite different from the one most institutional writers set for themselves. However, I have no urge to enlarge on this point. My intent here is not polemical, on behalf of satisficing postulates, but rather to show how large a plot of common ground is shared by optimizing and satisficing analysis. Again, compare Becker (1962).

with the characteristics of the objective environment in which they make their decisions. In such a world, we must give an account not only of *substantive rationality*—the extent to which appropriate courses of action are chosen—but also *procedural rationality*—the effectiveness, in light of human cognitive powers and limitations, of the *procedures* used to choose actions. As economics moves out toward situations of increasing cognitive complexity, it becomes increasingly concerned with the ability of actors to cope with the complexity, and hence with the procedural aspects of rationality. In the remainder of my talk, I would like to develop this concept of procedural rationality, and its implications for economic analysis.

III. Mind as the Scarce Resource

Until rather recently, such limited attention as was paid by economists to procedural, as distinct from substantive, rationality was mainly motivated by the problems of uncertainty and expectations. The simple notion of maximizing utility or profit could not be applied to situations where the optimum action depended on uncertain environmental events, or upon the actions of other rational agents (for example, imperfect competition).

The former difficulty was removed to some degree by replacing utility maximization with the maximization of subjective expected utility (*SEU*) as the criterion of rationality. In spite of its conceptual elegance, however, the *SEU* solution has some grave defects as either a normative or a descriptive formulation. In general, the optimal solution depends upon all of the moments of the frequency distributions of uncertain events. The exceptions are a small but important class of cases where the utility or profit function is quadratic and all constraints are in the form of equations rather than inequalities.[7] The empirical defect of the *SEU* formulation is that when it has been subjected to test in the laboratory or the real world, even in relatively simple situations, the behavior of human subjects has generally departed widely from it.

Some of the evidence has been surveyed by Ward Edwards, and more recently by Daniel Kahneman and Amos Tversky. They describe experimental situations in which estimates formed on the basis of initial information are not revised nearly as much by subsequent information as would be required by Bayes' Theorem. In other situations, subjects respond largely to the information received most recently, and take inadequate account of prior information.

Behavior that is radically inconsistent with the *SEU* framework occurs also in naturalistic settings. Howard Kunreuther et al. have recently carried out extensive studies of behavior and attitudes relating to the purchase of flood insurance by persons owning property in low-lying areas. They found that knowledge of the availability of insurance, or rates, and of objective risks was very imperfect, and that the actual decisions whether or not to insure were related much more to personal experience with floods than to any objective facts about the situation—or even to personal subjective beliefs about those facts. In the face of this evidence, it is hard to take *SEU* seriously as a theory of actual human behavior in the face of uncertainty.[8]

For situations where the rationality of an action depends upon what others (who are also striving to be rational) do again, no consensus has been reached as to what constitutes optimal behavior. This is one of the reasons I have elsewhere called imperfect competition "the permanent and ineradicable scandal of economic theory" (1976b, p. 140). The most imaginative and

[7]In this case the expected values of the environmental variables serve as certainty equivalents, so that *SEU* maximization requires only replacing the unknown true values by these expected values. See the author (1957).

[8]Kunreuther et al. point out that the theory cannot be "saved" by assuming utility to be radically nonlinear in money. In the flood insurance case, that interpretation of the data would work only if we were willing to assume that money has strongly *increasing* marginal utility, not a very plausible escape route for the theory.

ambitious attempt to resolve the difficulty was the von Neumann-Morgenstern theory of games, which is embarrassing in the wealth of alternative solutions it offers. While the theory of games reveals the potential richness of behavior when rational individuals are faced with conflict of interest, the capability of reacting to each other's actions (or expected actions), and possibilities for coalition, it has provided no unique and universally accepted criterion of rationality to generalize the *SEU* criterion and extend it to this broader range of situations.

The so-called "rational expectations" models, currently so popular (and due originally to Muth), pass over these problems rather than solving them. They ignore potential coalitions and attempted mutual outguessing behavior, and correspond to optimal solutions only when the losses are quadratic functions of the errors of estimate.[9] Hence they do not correspond to any classical criterion of rationality, and labeling them with that term, rather than the more neutral "consistent expectations," provides them with a rather unwarranted legitimation.

Finally, it should be remarked that the main motivation in economics for developing theories of uncertainty and mutual expectations has not been to replace substantive criteria of rationality with procedural criteria, but rather to find substantive criteria broad enough to extend the concept of rationality beyond the boundaries of static optimization under certainty. As with classical decision theory, the interest lies not in *how* decisions are made but in *what* decisions are made. (But see, contra, such analyses as Richard Cyert and Morris DeGroot.)

[9]That is, only under the conditions where the uncertainty equivalents of fn. 8 exist. Under other circumstances, a "rational" person would be well advised, if he knew that all others were following the "rational expectations" or "consistent expectations" rule, to recalculate his own optimal behavior on that assumption. Of course if others followed the same course, we would be back in the "outguessing" situation.

A. *Search and Teams*

Decision procedures have been treated more explicitly in the small bodies of work that have grown up in economics on the theory of search and on the theory of teams. Both these bodies of theory are specifically concerned with the limits on the ability of the economic actor to discover or compute what behavior is optimal for him. Both aspire not only to *take account* of human bounded rationality, but to *bring it within the compass* of the rational calculus. Let me explain what I mean by that distinction.

Problems of search arise when not all the alternatives of action are presented to the rational actor *ab initio*, but must be sought through some kind of costly activity. In general, an action will be chosen before the search has revealed all possible alternatives. One example of this kind of problem is the sale of a house, or some other asset, when offers are received sequentially and remain open for only a limited time (see the author, 1955). Another example which has been widely cited is the purchase of an automobile involving travel to dealers' lots (see Stigler, 1961). In both these examples, the question is not how the search is carried out, but how it is decided when to terminate it—that is, the amount of search. The question is answered by postulating a cost that increases with the total amount of search. In an optimizing model, the correct point of termination is found by equating the marginal cost of search with the (expected) marginal improvement in the set of alternatives. In a satisficing model, search terminates when the best offer exceeds an aspiration level that itself adjusts gradually to the value of the offers received so far. In both cases, search becomes just another factor of production, and investment in search is determined by the same marginal principle as investment in any other factor. However cavalierly these theories treat the actual search process, they do recognize explicitly that information gathering is not a free activity, and that unlimited amounts of it are not available.

The theory of teams, as developed by Marschak and Radner, goes a step farther in specifying the procedure of decision. That theory, as is well known, is concerned with the improvement that may be realized in a team's decisions by interchange of information among the team members. But here the theory does not limit itself to determining the aggregate amount of information that should be transmitted, but seeks to calculate what messages should be exchanged, under what conditions, and at what cost. The content of the communication as well as the total amount of information becomes relevant to the theory.

In its attitude toward rationality, the theory of teams is as "classical," however, as is search theory. The bounds on the rationality of the team members are "externalized" and represented as costs of communication, so that they can be folded into the economic calculation along with the costs and benefits of outcomes.

B. *Rational Search Procedures*

To find theories that compare the merits of alternative search procedures, we must look largely outside the domain of economics. A number of such theories have been developed in the past thirty years, mainly by management scientists and researchers in the field of artificial intelligence. An important example is the body of work that has been done on integer programming.

Integer programming problems resemble linear programming problems (to maximize some quantity, subject to constraints in the form of linear equations and inequalities), with the added condition that certain variables can only take whole numbers as their values. The integer constraint makes inapplicable most of the powerful computational methods available for solving linear programming problems, with the result that integer programming problems are far less tractable, computationally, than linear programming problems having comparable numbers of variables.

Solution methods for integer programming problems use various forms of highly selective search—for example branch-and-bound methods that establish successively narrower limits for the value of the optimum, and hence permit a corresponding narrowing of search to promising regions of the space. It becomes a matter of considerable practical and theoretical interest to evaluate the relative computational efficiency of competing search procedures, and also to estimate how the cost of search will grow with the size of the problem posed. Until recently, most evaluation of search algorithms has been empirical: they have been tested on sample problems. Recently, however, a body of theory—called theory of computational complexity—has grown up that begins to answer some of these questions in a more systematic way.

I cannot give here an account of the theory of computational complexity, or all of its implications for procedural rationality. A good introduction will be found in Alfred Aho et al. One important set of results that comes out of the theory does require at least brief mention. These results have to do with the way in which the amount of computation required to solve problems of a given class grows with the size of the problems—with the number of variables, say.[10]

In a domain where computational requirements grow rapidly with problem size, we will be able to solve only small problems; in domains where the requirements grow slowly, we will be able to solve much larger problems. The problems that the real world presents to us are generally enormous compared with the problems that we can solve on even our largest computers. Hence, our computational models are always rough approximations to the reality, and we must hope that the approximation will not be too inexact to be useful.

[10]Most of the theorems in computational complexity have to do with the "worst case," that is, with the maximum amount of computation required to solve *any* problem of the given class. Very few results are available for the expected cost, averaged over all problems of the class.

We will be particularly concerned that computational costs not increase rapidly with problem size.

It is customary in the theory of computational complexity to regard problems of a given size as "tractable" if computations do not grow faster than at some fixed power of problem size. Such classes of problems are known as "polynomial complex." Problems that grow exponentially in complexity with size are not polynomial complex, since the rate of growth of computation comes to exceed any fixed power of their size.

A large and important class of problems which includes the general integer programming problem, as well as standard scheduling problems, all have been shown to have the same level of complexity—if one is polynomial complex, then all are; if one is not polynomial complex, then none are. These problems have been labeled "*NP*-complete." It is conjectured, but not yet proven, that the class of *NP*-complete problems is not polynomially complex, but probably exponentially complex.

The significance of these findings and conjectures is in showing that computational difficulties, and the need to approximate, are not just a minor annoying feature of our world to be dealt with by manufacturing larger computers or breeding smarter people. Complexity is deep in the nature of things, and discovering tolerable approximation procedures and heuristics that permit huge spaces to be searched very selectively lies at the heart of intelligence, whether human or artificial. A theory of rationality that does not give an account of problem solving in the face of complexity is sadly incomplete. It is worse than incomplete; it can be seriously misleading by providing "solutions" to economic questions that are without operational significance.

One interesting and important direction of research in computational complexity lies in showing how the complexity of problems might be decreased by weakening the requirements for solution—by requiring solutions only to approximate the optimum, or by replacing an optimality criterion by a satisficing criterion. Results are still fragmentary, but it is already known that there are some cases where such modifications reduce exponential or *NP*-complete problem classes to polynomial-complete classes.

The theory of heuristic search, cultivated in artificial intelligence and information processing psychology, is concerned with devising or identifying search procedures that will permit systems of limited computational capacity to make complex decisions and solve difficult problems. (For a general survey of the theory, see Nils Nilsson.) When a task environment has patterned structure, so that solutions to a search problem are not scattered randomly throughout it, but are located in ways related to the structure, then an intelligent system capable of detecting the pattern can exploit it in order to search for solutions in a highly selective way.

One form, for example, of selective heuristic search, called best-first search, assigns to each node in the search space an estimate of the distance of that node from a solution. At each stage, the next increment of effort is expended in searching from the node, among those already reached, that has the smallest distance estimate (see, for example, the author and J.B. Kadane). As another example, when the task is to find a good or best solution, it may be possible to assign upper and lower bounds on the values of the solutions that can be obtained by searching a particular part of the space. If the upper bound on region *A* is lower than the lower bound on some other region, then region *A* does not need to be searched at all.

I will leave the topics of computational complexity and heuristic search with these sketchy remarks. What implications these developments in the theory of procedural rationality will have for economics defined as "the science which treats of the wealth-getting and wealth-using activities of man" remain to be seen. That they are an integral part of economics defined as "the science which treats of the allocation of scarce resources" is obvious. The scarce resource is computational capacity—the mind. The ability of man to solve complex problems,

and the magnitude of the resources that have to be allocated to solving them, depend on the efficiency with which this resource, mind, is deployed.

C. *Attention as the Scarce Resource*

Finally, I would like to turn from the rather highly developed approaches to procedural rationality that I have been discussing back to the more qualitative kinds of institutional issues that were considered in the previous section of this paper. Many of the central issues of our time are questions of how we use limited information and limited computational capacity to deal with enormous problems whose shape we barely grasp.

For many purposes, a modern government can be regarded as a parallel computing device. While one part of its capability for rational problem solving is directed to fire protection, another is directed to paving highways, and another to collecting refuse. For other important purposes, a government, like a human being, is a serial processing system, capable of attending to only one thing at a time. When important new policies must be formulated, public and official attention must be focused on one or a few matters. Other concerns, no matter how pressing, must wait their turn on the agenda. When the agenda becomes crowded, public life begins to appear more and more as a succession of crises. When problems become interrelated, as energy and pollution problems have become, there is the constant danger that attention directed to a single facet of the web will spawn solutions that disregard vital consequences for the other facets. When oil is scarce, we return to coal, but forget that we must then deal with vastly increased quantities of sulfur oxides in our urban air. Or we outlaw nuclear power stations because of radiation hazards, but fail to make alternative provision to meet our energy needs. It is futile to talk of substantive rationality in public affairs without considering what procedural means are available to order issues on the public agenda in a rational way, and to insure attention to the indirect consequences of actions taken to reach specific goals or solve specific problems.

In a world where information is relatively scarce, and where problems for decision are few and simple, information is almost always a positive good. In a world where attention is a major scarce resource, information may be an expensive luxury, for it may turn our attention from what is important to what is unimportant. We cannot afford to attend to information simply because it is there. I am not aware that there has been any systematic development of a theory of information and communication that treats attention rather than information as the scarce resource.[11] Some of the practical consequences of attention scarcity have already been noticed in business and government, where early designs of so-called "management information systems" flooded executives with trivial data and, until they learned to ignore them, distracted their attention from more important matters. It is probably true of contemporary organizations that an automated information system that does not consume and digest vastly more information than it produces and distributes harms the performance of the organization in which it is incorporated.

The management of attention and tracing indirect consequences of action are two of the basic issues of procedural rationality that confront a modern society. There are others of comparable importance: what decision-making procedure is rational when the basic quantities for making marginal comparisons are simply not known? A few years ago, I served as chairman of a National Academy of Sciences (*NAS*) committee whose job it was to advise the Congress on the control of automobile emissions (see *NAS*, Coordinating Committee on Air Quality Studies). It is easy to formulate an *SEU* model to conceptualize the problem. There is a production function for automobiles that associates different costs with different levels of emissions. The laws govern-

[11]Some unsystematic remarks on the subject will be found in the author (1976a, chs. 13, 14).

ing the chemistry of the atmosphere determine the concentrations of polluting substances in the air as a function of the levels of emissions. Biomedical science tells us what effects on life and health can be expected from various concentrations of pollutants. All we need do is to attach a price tag to life and health, and we can calculate the optimum level of pollution control.

There is only one hitch—which will be apparent to all of you. None of the relevant parameters of the various "production functions" are known—except, within half an order of magnitude, the cost of reducing the emissions themselves. The physics and chemistry of the atmosphere presents a series of unsolved problems—particularly relating to the photochemical reactions affecting the oxides of nitrogen and ozone. Medical science is barely able to detect that there *are* health effects from pollutants, much less measure how large these effects are. The committee's deliberations led immediately to one conclusion—one that congressmen are accustomed to hearing from such committees: We need more research. But while the research is being done, what provisions should be incorporated in the Clean Air Act of 1977 (or the Acts of 1978 through 2000, for that matter)? For research won't give us clear answers then either. What constitutes procedural rationality in such circumstances?

"Reasonable men" reach "reasonable" conclusions in circumstances where they have no prospect of applying classical models of substantive rationality. We know only imperfectly how they do it. We know even less whether the procedures they use in place of the inapplicable models have any merit—although most of us would choose them in preference to drawing lots. The study of procedural rationality in circumstances where attention is scarce, where problems are immensely complex, and where crucial information is absent presents a host of challenging and fundamental research problems to anyone who is interested in the rational allocation of scarce resources.

IV. Conclusion

In histories of human civilization, the invention of writing and the invention of printing are always treated as key events. Perhaps in future histories the invention of electrical communication and the invention of the computer will receive comparable emphasis. What all of these developments have in common, and what makes them so important, is that they represent basic changes in man's equipment for making rational choices—in his computational capabilities. Problems that are impossible to handle with the head alone (multiplying large numbers together, for example) become trivial when they can be written down on paper. Interactions of energy and environment that almost defy conceptualization lend themselves to at least approximate modeling with modern computers.

The advances in man's capacity for procedural rationality are not limited to these obvious examples. The invention of algebra, of analytic geometry, of the calculus were such advances. So was the invention, if we may call it that, of the modern organization, which greatly increased man's capacity for coordinated parallel activity. Changes in the production function for information and decisions are central to any account of changes over the centuries of the human condition.

In the past, economics has largely ignored the processes that rational man uses in reaching his resource allocation decisions. This was possibly an acceptable strategy for explaining rational decision in static, relatively simple problem situations where it might be assumed that additional computational time or power could not change the outcome. The strategy does not work, however, when we are seeking to explain the decision maker's behavior in complex, dynamic circumstances that involve a great deal of uncertainty, and that make severe demands upon his attention.

As economics acquires aspirations to explain behavior under these typical conditions of modern organizational and public life, it will have to devote major energy to

building a theory of procedural rationality to complement existing theories of substantive rationality. Some elements of such a theory can be borrowed from the neighboring disciplines of operations research, artificial intelligence, and cognitive psychology; but an enormous job remains to be done to extend this work and to apply it to specifically economic problems.

Jacob Marschak, throughout his long career, had a deep belief in and commitment to the interdependencies and complementarity of the several social sciences. I have shared that belief and commitment, without always agreeing with him in detail as to the precise route for exploiting it. The developments I have been describing strengthen greatly, it seems to me, the rational grounds for both belief and commitment. Whether we accept the more restricted definition of economics that I quoted from Ely's textbook, or the wider definition that is widely accepted today, we have every reason to try to communicate with the other social sciences, both to find out what we have to say that may be of interest to them, and to discover what they can teach us about the nature of procedural rationality.

REFERENCES

Alfred V. Aho et al., *The Design and Analysis of Computer Algorithms*, Reading 1974.

Chester I. Barnard, *The Functions of the Executive*, Cambridge 1938.

G. S. Becker, "Irrational Behavior and Economic Theory," *J. Polit. Econ.*, Feb. 1962, *70*, 1-13.

———, "A Theory of Social Interations," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1063-93.

F. M. Cancian, "Functional Analysis," in *International Encyclopedia of the Social Sciences*, 1968, *6*, 29-42.

R. M. Cyert and M. H. Degrott, "Sequential Strategies in Dual Control," *Theory Decn.*, Apr. 1977, *8*, 173-92.

Anthony Downs, *An Economic Theory of Democracy*, New York 1957.

Maurice Duverger, *Political Parties*, rev. ed., New York 1959, (*Les Partis Politiques*, Paris 1951).

W. Edwards, "Conservation in Human Information Processing," in Benjamin Kleinmuntz, ed., *Formal Representation of Human Thought*, New York 1968.

Richard T. Ely, *Outlines of Economics*, rev. ed., New York 1930.

S. Freud, "Five Lectures on Psychoanalysis" (originally "The Origin and Development of Psychoanalysis" 1910) in *The Complete Psychological Works of Sigmund Freud*, Vol. 11, London 1957.

George Homans, *Social Behavior: Its Elementary Forms*, New York 1961.

D. Kahneman and A. Tversky, "On the Psychology of Prediction," *Psychol. Rev.*, July 1973, *80*, 237-51.

Janos Kornai, *Anti-Equilibrium*, Amsterdam 1971.

Howard Kunreuther et al., *Protecting Against High-Risk Hazards: Public Policy Lessons*, New York 1978.

James G. March and Herbert A. Simon, *Organizations*, New York 1958.

Jacob Marschak and Roy Radner, *Economic Theory of Teams*, New Haven 1972.

John M. Montias, *The Structure of Economic Systems*, New Haven 1976.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315-35.

Nils Nilsson, *Problem-Solving Methods in Artificial Intelligence*, New York 1971.

A. Rees, "Economics," in *International Encyclopedia of the Social Sciences*, 1968, *4*, 472.

William H. Riker, *The Theory of Political Coalitions*, New Haven 1962.

——— **and Peter C. Ordeshook,** *An Introduction to Positive Political Theory*, New Jersey 1973.

Paul Samuelson, *Foundations of Economic Analysis*, Cambridge 1947.

George Simmel, *Soziologie*, Berlin 1908.

Herbert A. Simon, "A Formal Theory of the Employment Relation," *Enconometrica*, July 1951, *19*, 293-305.

———, "A Behavioral Model of Rational

Choice," *Quart. J. Econ.*, Feb. 1955, *69*, 99–118.

———, "Dynamic Programming Under Uncertainty with a Quadratic Criterion Function," *Econometrica*, Jan. 1956, *24*, 74–81.

———, (1976a) *Administrative Behavior*, 3d ed., New York 1976.

———, (1976b) "From Substantive to Procedural Rationality," in Spiro J. Latsis, ed., *Method and Appraisal in Economics*, Cambridge 1976.

——— **and J. B. Kadane**, "Optimal Problem-Solving Search: All-or-None Solutions," *Artificial Intel.*, Fall 1975, *6*, 235–48.

G. J. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–15.

——— **and G. S. Becker**, "De Gustibus non est Disputandum," *Amer. Econ. Rev.*, Mar. 1977, *67*, 76–90.

Oliver E. Williamson, *Markets and Hierarchies*, New York 1975.

National Academy of Sciences, (*NAS*) Coordinating Committee on Air Quality Studies, *Air Quality and Automobile Emission Control*, Vol. 1 summary rep., Washington 1974.

What do Economics Majors Learn?

By DAVID G. HARTMAN*

The introductory economics course has been studied extensively, with respect to both course content and instructional methods. Far less is known about the remainder of the program for undergraduate economics majors. As an initial step toward understanding and improving the educational experience of economics majors, this paper examines what economics majors learn. Specifically, the primary results give information about what students at Harvard learn, but it is hoped that insights into the experience of students in general can be gained. Because my primary concern is with general economic understanding, the emphasis of this paper will be on how well students learn micro- and macroeconomic concepts and how to apply them. The study begins with a few casual observations.

Before any formal analysis was undertaken, there were reasons to believe that improvements in the economics program were needed. First, as one who has taught introductory economics at Harvard and also been in charge of the general examinations required of all graduating economics majors, I have often suspected that either students in the introductory course know much less than it appears or there is a sizable group of economics majors who leave school with little more than the level of economic understanding they achieved by the end of their first course. This evidence is, of course, of limited usefulness because the passage of time would tend to erode the students' skills. Moreover, most graduating seniors are at least a year away from their last general micro or macro course. A more formal analysis will be presented below to isolate the impacts of various parts of the economics program on student performance.

A secondary source of casual evidence is the students themselves. It is no secret that a large majority of Harvard economics undergraduates feel that the introductory course is the high point of the program, with a number of the subsequent courses, particularly the intermediate theory sequence, considered highly repetitious of material covered in introductory economics. The feeling is widespread that a general analytical ability is not being developed as it should be in an undergraduate program. There are several points to be considered about such student complaints. First, having witnessed similar students feelings at another school (where the point was not made as vocally as at Harvard), I find it difficult to accept this dissatisfaction as unique. Whether it highlights a significant problem is a separate issue. Although I tend to take fairly seriously the complaint of a broad group of students that courses are not sufficiently rigorous, it is certainly possible that a great deal of useful skill is being acquired. It may be that some courses only seem repetitious because the beginning course provides at least a brief introduction to all the major topics to be examined. Graduate teaching assistants in the intermediate micro-macro sequence support this hypothesis by pointing out that students often perform poorly because, having the illusion of already knowing the material, they put too little effort into the intermediate courses. Finally, it is far from clear, without further investigation, that repetition is not a valuable use of time, even if students complain.

*Harvard University. I wish to thank President Derek C. Bok for his interest in and funding of this effort. I owe Elizabeth Allison a great debt for her encouragement and advice. Liam P. Ebrill, Jeff Wolcowitz, and Ken Sokoloff did a heroic job of extracting the needed data from general examinations. David Lindauer provided many useful comments on an earlier draft.

TABLE 1—CHARACTERISTICS OF HONORS AND NONHONORS ECONOMICS STUDENTS

	Honors Students	Nonhonors Students	All Students
Number in entire 1977 class	61	118	179
Number of students in sample	50	44	94
Number of students having taken micro courses:			
Traditional micro course	43	26	69
Policy micro course	4	5	9
Graduate micro course	3	2	5
Number of students having taken no micro course	0	11	11
Number of students having taken macro courses:			
Traditional macro course	27	18	45
Graduate macro course	23	10	33
Number of students having taken no macro course	0	16	16
Average introductory micro course grade (0-15 scale)	11.45	9.98	10.79
Average introductory macro course grade (0-15 scale)	11.86	10.56	11.27

These issues are essentially empirical; empirical evidence to be presented will hopefully provide some answers. The investigation will proceed by first discussing specific goals of the economics program, then suggesting a method for measuring how well students meet these goals, and finally attempting to discover the contributions of various parts of the program to the students' achievement.

The general goals of an economics program are highly controversial, as Robert Horton and Dennis Weidenaar discovered in their survey. The more specific objectives implicit in the Harvard method of evaluating graduating students seem roughly consistent with the Horton and Weidenaar "consensus" goal; the Harvard general examinations in economics are designed to test a student's ability to use economic tools to answer real world questions. Specifically, a student's knowledge of macro theory, knowledge of micro theory, ability to apply macro theory, and ability to apply micro theory are evaluated and can be given separate scores. For purposes of this paper, attainment of skill in those four categories will be taken as the objective of the program. Because it represents the best information available, the score on the general exam in each category will be used to measure what an economics student has learned. The problems associated with using test scores to represent achievement should be kept in mind as the empirical results are analyzed.

I. Majoring in Economics at Harvard

Economics students choose between an "honors" major and a "nonhonors" major. In a typical year, there are about 75 honors and 125 nonhonors majors. Only honors students are required to complete a course in both intermediate micro and macro and to write a senior thesis. Although nonhonors students have neither requirement, many complete the intermediate theory courses because of an interest in the topic, because they wish to be well prepared for the general exam, or because they had at some time intended to pursue an honors program. A good grade record is required for graduation with honors, so there is a natural presumption that honors majors are better students. However, there are a number of reasons for good students to do a nonhonors program, including the desire to take a wider range of courses than the honors requirements would allow. Table 1

gives comparisons of 1977 honors and nonhonors students in terms both of grades in introductory courses and of elective courses taken. The comparisons are based on the sample of students to be used in the empirical study.[1]

Honors students have a number of options in meeting the intermediate micro-macro requirement. The courses with the largest numbers of students are the "traditional" micro and macro offerings. The empirical investigation which follows is based on students who took the 1977 general examinations, but who took the traditional micro or macro course at a variety of points in their careers, with different instructors and texts. In addition, the traditional courses apparently are becoming more like the "alternative" intermediate courses than were the traditional courses taken by most 1977 graduates.

There are two other courses in micro which satisfy the honors requirement. Almost 10 percent of the sample students took the policy and applications oriented micro course, which is taught by the case method with a large number of problem sets and a great deal of assigned reading. The other course, with about 5 percent of the sample students, is taught at nearly the level of sophistication of the course intended for economics Ph.D. students. It requires a mathematical preparation beyond most undergraduates and is also taken by graduate students from other departments. Because of this variety in the Harvard program, it will be possible to determine the impact on learning of the traditional micro course compared to two very different alternatives (to be called the "policy" micro course and the "graduate" micro course); a reference group is available since 12 percent of the students had had no micro course past the introductory level.

There is only one alternative to the traditional macro offering. It was taken by 35 percent of the sample students. Despite the fact that this course is intended to be taught at nearly the level of the course for economics Ph.D. students, and is taken by graduate students from other departments, the level of math required is not above that of most Harvard undergraduates. The material presented not only is more difficult than that in the traditional macro course (particularly since the course employs an extensive reading list rather than a basic text), but also puts more emphasis on policy. An attempt will be made to assess the impacts of the traditional course and this graduate course on student performance. Once again, the reference group consists of those students (17 percent of the sample) who had had neither course.

II. Measurement and Empirical Results

Graders of the 1977 general examinations were asked to assign each student four grades (each on a 0–10 scale): 1) micro theory; 2) micro application; 3) macro theory; and 4) macro application. Each question was graded by two graduate students familiar with the undergraduate micro-macro courses. Since different questions are asked on the honors and nonhonors exams, it was important that specific grading standards be adopted. The graders reported little problem with consistency in scoring the two exams because of the similarity and generality of the questions asked. However, a concern was expressed, particularly by the macro graders, about their ability to separate the students' knowledge of theory from their ability to apply it, based on answers to these general questions. The empirical results tend to confirm this difficulty with macro; I suspect that it is in the nature of macroeconomics for theory and application to be so closely related.

To allow a direct confirmation of grading consistency across the two examinations, a "practice examination" was given to thirty of the students two weeks prior to the general exams. The practice test consisted of multiple choice questions chosen to fall into the four categories of interest. By hav-

[1]The 94-student sample on which these comparisons are based results from requiring that personal data be available and that the students included took the spring 1977 exams. Nonhonors students may substitute a fall exam, so they are underrepresented in the sample. I have no reason to believe, however, that this sample is biased in any way critical to the results.

ing a set of scores on a common test to compare with the scores on the honors and nonhonors exams, it was possible to verify that in no case was there any significant bias in the scoring of honors and nonhonors general exams.

Separate regressions were run to estimate the impact of courses taken as part of an economics major on students' measured knowledge of micro theory, ability to apply micro theory, knowledge of macro theory, and ability to apply macro theory. A simple model is assumed: an economics major's level of skill in each of the four categories depends on his/her innate ability, the skill developed in introductory economics, and the amount learned in the relevant courses taken as part of the economics program. Measurement of the first two factors will be discussed as the empirical results are presented. The contribution of the economics program to a student's skill, obviously the factor most important to isolate for purposes of this paper, is estimated using a set of variables indicating which courses a student has taken. With respect to microeconomics, three dummy variables indicate whether a student has taken the traditional micro course, the policy micro course, or the graduate micro course (or, of course, none of these). Another variable is the number of "micro related field" courses a student has taken.[2] In the macroeconomics regressions, two dummy variables indicate whether a student has taken the traditional macro course or the graduate macro course (or neither). Another variable is the number of macro related field courses a student has taken.[3]

In the initial attempt to explain students' knowledge of micro theory (equation (1) in Table 2), grades in the microeconomics half of the introductory course were used to represent both ability and the learning acquired in the course. The rest of the equation consists of the "economics major" variables discussed above. The dependent variables and introductory course grade variables are expressed in *logs* in every estimation. So, the coefficients on the micro course dummy variables and the field course variable in (1) are the estimated percentage increases in micro theory score produced by taking the associated course. Therefore, the traditional course is estimated to increase one's micro theory score by about 24 percent, the policy micro course by 44 percent, the graduate micro course by 51 percent, and an additional micro field course by 9 percent. All of these effects are significant at or very close to significance at the 5 percent level.

These results should, however, be regarded with suspicion because of inclusion of only the introductory course grade to measure ability. In particular, the equation (1) result could occur as a consequence of students choosing courses based on their ability. For example, if the most able economics students choose the graduate micro course, its larger estimated impact need not be evidence of any learning premium over the traditional course. I would anticipate that the most significant bias results from not having any measure of math skill in equation (1), since the best information a student has about his aptitude for micro (aside from math) is probably his introductory course grade.

Numerous measures of student ability including SAT scores, high school class rank (adjusted for school size), and interview ratings, as well as data on race and sex, were available from admissions information. The SAT math scores, as expected, were the only significant ability factor in any regression. The inclusion of the SAT scores (equation (2)) reduces to insignificance the introductory micro course grade variable. Since the sign is then wrong, the coefficient is constrained to be zero.

[2]As used in this study, microeconomics related field courses are defined as: economic principles and public policy; public finance; development economics; international trade and investment; economics of managerial decisions; business organization and behavior; markets and market structure; labor economics; urban economics.

[3]As used in this study, macroeconomics related field courses are defined as: monetary theory and financial institutions; applied macroeconomics; public finance; international [illegible] economics; development economics.

The resulting equation (3) has substantially lower estimated impacts for components of the economics program, confirming the extent of bias introduced when math ability is left out. Now, only the policy micro course meets the strict criterion for significance normally employed; the number of micro field courses does, however, have a significant impact at the .05 level in a one-tail test. The graduate micro course (.10 significance level) and the traditional course (.20 level) do not fare so well. The small number of students having taken the graduate course (see Table 1) may be a contributing factor to its insignificance.

The corresponding equation ((6)) for ability to apply micro concepts indicates that, of the components of the economics major, only the policy micro course has an impact on student performance significant at even the .10 level. The traditional course and the field courses produce an estimated improvement in a student's ability to apply microeconomics with a significance of .20; the influence of the graduate micro course is even less significant.

The results explaining students' knowledge of macro theory and their ability to apply it lend suppon to the contention that it is difficult to separate the two scores: the results are very similar. The results also show that math ability does not play a significant independent role.

Table 2 shows the significance level of the graduate macro course in explaining the macro theory score on the general examinations is .001, while the traditional course and the macro field courses are significant at the .05 level. Taking the graduate macro course produces an estimated 21 percent increase in exam score, while the impact of the traditional course is 11 percent and the effect of each macro field course is an estimated 4 percent.

The large differences in the macro theory and applications scores seem to imply that the traditional course and the field courses have a less significant effect on students' ability to apply macro concepts. It should also be noted that the R^2 of the applications equation is substantially below that of the macro theory equation. This implies that the ability to apply macro concepts is more difficult to measure and/or is determined more by the ability factors not controlled for, than is knowledge of macro theory.

TABLE 2

Equation	Dependent Variable (*logs*)	Constant	Introductory Course Grade (*logs*)		Traditional Intermediate Course (dummy)		Policy Intermediate Course (dummy)	Graduate Intermediate Course (dummy)		Field Courses Taken (number)		SAT Math Score	SAT Verbal Score	R^2
			Micro	Macro	Micro	Macro	Micro	Micro	Macro	Micro	Macro			
(1)	Micro Theory Score	1 11	.088		.239		.438	.510		.094				.14
		(4.12)	(.83)		(1.92)		(2.47)	(2.36)		(2.69)				
(2)		−7.21	− 063		.129		360	302		.056		1.263	.096	.31
		(3.39)	(.60)		(1.09)		(2.16)	(1.46)		(1.67)		(2.82)	(.25)	
(3)		−6.91			.126		.360	.300		.057		1.177	.112	.31
		(3.36)			(1.07)		(2.17)	(1.46)		(1.71)		(2.79)	(.30)	
(4)	Micro Application Score	1.31	−.018		.303		.375	384		.096				.06
		(3.92)	(.14)		(1.97)		(1.71)	(1.44)		(2.22)				
(5)		−9.26	−.222		.168		.294	119		.047		1.859	−.133	.27
		(3.58)	(1 73)		(1.17)		(1.44)	(.47)		(1.15)		(3.41)	(.29)	
(6)		−8.20			.154		.294	.114		.050		1.559	−.077	.24
		(3.22)			(1.06)		(1.43)	(.45)		(1.21)		(2.98)	(.17)	
(7)	Macro Theory Score	1.34		.174		107			.218		.039			.37
		(7.03)		(2.13)		(1.90)			(3.32)		(1.78)			
(8)		1.22		.173		107			.214		.040	.103	−.085	.38
		(1.07)		(1.98)		(1.87)			(3.03)		(1.80)	(.46)	(.43)	
(9)	Macro Application Score	1.58		.073		.086			.231		.040			.28
		(6.91)		(.74)		(1.27)			(2.93)		(1.54)			
(10)		.78		.054		.081			.213		.042	.129	.002	.29
		(.57)		(.51)		(1.17)			(2.52)		(1.58)	(.48)	(.01)	

Note: Absolute values of t-statistics are shown in parentheses. To obtain all data, sample was restricted to 94 students. Therefore, significance levels (one-tail test) are approximately: $t = 2.37$, significance level = .01; $t = 1.99$, significance level = .025; $t = 1.66$, significance level = .05; $t = 1.29$, significance level = .10.

III. Summary and Conclusions

There are numerous reasons for statistically insignificant results and, therefore, there is great risk in placing too much emphasis on the empirical evidence presented here. However, those concerned with what economics majors learn can take little comfort from this analysis.

The traditional microeconomics course appears to have a small impact on either students' knowledge of micro theory or their ability to apply it to real world problems at the end of their college careers. One of the alternatives offered to Harvard students, a graduate-type highly theoretical course, does a bit better in adding to the students' knowledge of micro theory, but is worse when it comes to teaching them to apply the theory. Only the policy oriented micro course has a verifiable effect on the students' understanding of micro theory; even so, its impact on their ability to apply micro concepts is below usually acceptable significance levels, once adjustments are made for mathematical ability.

The traditional intermediate macro course gives evidence of improving macro theory scores, although its effect on ability to apply those concepts is below usual significance standards. The graduate level macro course, which is oriented to policy and not particularly mathematical although demanding, is a highly significant explainer of student knowledge. Finally, the micro and macro field courses tend overall to have a moderate impact on student scores.

At the risk of overemphasizing these admittedly rough statistical conclusions, it appears that there is a problem with the economics major. After years of study and improvement in the beginning course, it seems time to begin a similar effort with respect to the rest of the undergraduate program. It is not surprising that with a beginning course which has become a thorough and rigorous introduction to the discipline, students find the traditional courses which follow repetitious. They may be mistaken and, in any event, repetition may be quite valuable, but the evidence available does not substantially dispute their conclusions. Study and innovation, which are taking place in traditional courses at Harvard in response to the changes in the knowledge possessed by students emerging from the introductory class, must be encouraged. At a minimum, new methods of presenting information should be devised to prevent student dissatisfaction. From Harvard's experience with alternative courses, it would appear that at least the average student is capable of learning significantly more difficult material, by reading more extensively and working on more independent assignments, than is normally taught in intermediate courses. On the other hand, it seems unproductive to offer courses too much like those given for graduate students because, without experience in problem solving, undergraduates do not learn to apply the theory they are taught.

As indicated at the start, this paper represents just an initial step. The conclusions are necessarily imprecise and must be sharpened by study of systems at other schools. Until such broader information is available, application of these results to other than the specific courses studied here is quite risky. The effects on student learning of courses other than those in micro and macro and the impact of the economics program on knowledge retained years after graduation are also areas in which further research is necessary.

REFERENCES

R. V. Horton and D. J. Weidenaar, "Wherefore Economic Education?," *J. Econ. Educ.*, Fall 1975, *7*, 40–44.

ECONOMICS AND ANTHROPOLOGY: DEVELOPING AND PRIMITIVE ECONOMIES

Is Economic Anthropology of Interest to Economists?

By GEORGE DALTON*

The answer to the question posed in the title of this paper is yes, but only to the few who study preindustrial economic history,[1] the few who work on agricultural development of rural communities in the Third World today,[2] and the few interested in institutionalist themes of political economy.[3] To explain why this is so requires a description of economic anthropology which contrasts it with economics.

Economics is a large subject and an international subject, heavily theoretical, which makes extensive use of statistical series, econometrics, and other sorts of applied mathematics. Overwhelmingly, its empirical focus of interest has been the dozen or so industrial capitalist countries, that is, the national market economies of Britain, Western Europe, and America in the 200 years since industrialization began. From the Physiocrats, Mercantilists, and Adam Smith to the present, economics has always had strong policy concerns: to measure and analyze in order to understand how to improve economic performance, for example, to counteract depression and inflation and to accelerate rates of growth. The few fields of economics which do not fit this description are either small fields (preindustrial economic history, comparative economic systems, i.e., Soviet and other communist economies) or recent fields (economic development of the Third World). Marxian economics, which we have had with us since the times of John Stuart Mill and Stanley Jevons has never seriously encroached on the prevailing paradigms initiated by David Ricardo, Alfred Marshall, John Maynard Keynes, and Paul Samuelson.

Economic anthropology is utterly different. It is a tiny subject having relatively few specialists compared to the numbers of anthropologists who specialize in politics, kinship, religion, language, myth, etc.[4] There are bits of quantification,[5] but mathematics and measurement are still occasional intrusions, not an integral part of the subject. Its empirical focus of interest is the village-level economy and (where present) the kingdom-states of Africa, Asia, Latin America, the Middle East, and islands of the Pacific Ocean: bands, tribes, and peasantries, whose preindustrial economies and societies are very different from those of modern industrial Europe and America. Most of the available basic descriptive information has been collected in the course of fieldwork observation by individual anthropologists who lived in such small communities for a year to two and wrote up what they observed in ethnographies.[6] Usually this means one anthropologist aided by a few local assistants. Team research is still uncommon.

*Departments of economics and anthropology, Northwestern University. The numerous references cited in this paper may be found in the three publications cited in the references.

[1]See Philip Grierson; M. I. Finley; Thomas C. Smith.

[2]See Irma Adelman and Cynthia Taft Morris (1967, especially chs. 5 and 8); Gunnar Myrdal (1957); John De Wilde; Clifton Wharton; Tariq Husain.

[3]See Karl Polanyi (1947); Myrdal (1960); Richard M. Titmuss.

[4]For example, see the areas of specialization anthropologists list for themselves in the American Anthropological Association.

[5]See Adelman and the author (1971a, b); Stuart Plattner (1975).

[6]Some good examples are Raymond Firth (1939) and T. Scarlett Epstein (1962).

Analysis in order to make policy to improve economic performance has not been a major concern of anthropologists.[7] Until recently only a trifling few anthropologists have worked on village-level development. So far we have only the beginnings of a theory of micro development, nothing like a confident ability to make policy pronouncements based on firm theoretical understanding.[8] Finally, Marxism has recently come to have an important influence in anthropology.

There are three characteristics of economic anthropology which quite directly shape the subject as it presently exists: extreme diversity among the very large set of small economies to be analyzed; extreme diversity in theoretical approach, that is, the absence of a prevailing paradigm of unified theory widely shared; and a recent growth of interest in economic anthropology, by anthropologists as well as persons in several social sciences and several branches of history. Almost all the economies traditionally studied by anthropologists are now studied by others as well.

An economist who comes to work in economic development is struck by the diversity among the hundred national economies we call less developed countries—India is very different from Liberia, Brazil from Upper Volta, Taiwan from Uganda—and by the importance to their development of the special institutional configuration conferred on each one of these countries by its history, politics, colonial experience, and ethnic composition. It is these cultural, historical, and political diversities, these institutional idiosyncrasies, and the markedly different levels of development achieved so far among these less developed countries which have forced some development economists to contrive unconventional analyses under such headings as "dualism" (see W. Arthur Lewis), "cumulative causation" (see Myrdal, 1957), "growth without development" (see R. W. Clower et al.), "the limitations of the special case" (see Dudley Seers); and to contrive unorthodox statistical, qualitative, and interdisciplinary techniques, as with the analyses of variance employed by Adelman and Morris (1967, 1973, 1978), and the psychological approach to entrepreneurship employed by Everett Hagen.

An economist who comes to work in economic anthropology is staggered by even greater diversity. India, a single large country in The Third World, contains half a million village communities which differ from one another in ways which are both specifiable and related directly to their economic performance.[9] Moreover, in order to comprehend the full set of economies which interests anthropologists, we must add to the millions of villages in the economist's set of today's hundred underdeveloped nations, such historical kingdoms as the Inca before the Spanish conquest and the Bugunda of East Africa before British colonial rule; and such tiny, stateless economies as the bands and clan segments of Australian aborigines, Eskimos, North American Indians, Highland New Guinea peoples, and Indians in the South American rainforest. All these economies are studied as they were organized before and after they were colonized and after the nations that contain them achieved political independence. Just as economic historians study the national economies of England, France, and Japan, and their component sectors and social groups throughout the 2,000 years or more of recorded history for those countries, so too do anthropologists study villages and kingdoms as they were structured at different historical time periods.

Most economic anthropology in print analyzes the small communities in which

[7]There is a small field called applied anthropology which for the most part concerns itself with successes and failures experienced by visiting experts attempting to introduce modern innovations piecemeal (usually modern agricultural or health practices) into underdeveloped, hinterland villages, a field of use to agricultural extension agents and persons in the Peace Corps. See the journal *Human Organization;* also, Edward H. Spicer (1952) and the author (1971a).

[8]For attempts at policy suggestions for rural development, see Epstein (1973); Adelman and the author (1971a,b); Husain.

[9]See Adelman and the author (1971 a,b).

anthropologists have done fieldwork over the last sixty years, beginning with Malinowski's pioneer work on islands off the East Coast of New Guniea. However, it is quite usual for anthropologists to write historical accounts of early economies based on documentary evidence and oral tradition, yielding descriptions of the Inca economy as it was organized just before the Spanish conquest of 1532,[10] and of the Tio Kingdom in Central Africa in the 1880's just before European colonial rule (see Jan Vansina). Anthropologists also have written historical accounts of economic change brought by European colonial rule, from 1500 to 1960, for example, in Indonesia (see Clifford Geertz), East Africa (see Audrey Richards et al.), and Latin American and the Caribbean.[11] To these we must now add the plethora of writings by economic and social historians of Third World countries, subjects that have grown sufficiently large to call forth specialized journals such as *African Economic History*, and interdisciplinary journals such as *The Journal of Peasant Studies*.

From the 1950's onward, changes in the real world together with changes in several university subjects have deepened and widened interest in the economies in which anthropologists have traditionally centered their fieldwork. These changes have also intensified disputes over which of three contending theoretical frameworks provides the most useful concepts and the most persuasive explanations of the structure of preindustrial economies and of their development. The changes in the real world I refer to are of course the ending of colonial rule and the establishment of some seventy new nation-states in Africa, Asia, and elsewhere to join—as underdeveloped countries—those of southern Europe and those of Latin America and the Caribbean which had achieved political independence a hundred or more years earlier.

What has also happened recently is that archaeology and several branches of preindustrial history have intensified their interest in economic organization, economic change, and particularly in the nature of early foreign trade, early usages of monetary objects, and the economic organization of early kingdom-states. The greatest growth in research and publications has occurred in historical subjects which either were very small twenty years ago, such as the economic history of Japan and China and the economic history of colonialism, or did not then exist at all, such as the economic history of Africa. A convergence of interests has taken place in the sense that there are now a dozen fields of social science and history which, like economic anthropology, are concerned with all sorts of preindustrial economies and their modern transformation.

This growth of interest in economic anthropology, economic archaeology, and the economic history of Third World countries has been accompanied by the emergence of three rival theoretical frameworks, all employed to analyze preindustrial economies: "formalism";[12] "substantivism";[13] and Marxism.[14]

To conclude: what kinds of questions does economic anthropology help to answer? What has economic anthropology achieved so far? What are promising lines of research?

Economic anthropology provides factual information, conceptual categories, and analytical insights of the same sorts provided by several of the subjects listed in Table 1, particularly preindustrial economic history (including the economic history of colonialism), rural sociology, agricultural history, rural development, and modernization of the Third World.

Evidence is accumulating from historical and anthropological research to support the conclusion that a number of economic institutions were originally contrived by early states to serve governmental rather than commercial purposes, particularly coined

[10]See John V. Murra; Sally Falk Moore.

[11]See Eric Wolf, chs. 8–11; Sidney Mintz.

[12]See C. Scott Cook; Harold K. Schneider.

[13]See Polanyi (1957); Marshall Sahlins; the author (1975).

[14]See Maurice Godelier; Emanuel Terray; Y. I. Semenov; Maurice Bloch.

TABLE 1—OTHER SUBJECTS DEALING WITH ECONOMY AND SOCIETY

Preindustrial Economies and Societies	Archaeology (Renfrew, 1972) Classical Antiquity (Finley, 1973) History of Preindustrial Europe (Bloch, 1967) History of Preindustrial Japan, China, Latin America, Africa, Middle-East (Smith, 1959; Fairbank, 1968; Gibson, 1964; Wilks, 1975
	↑ Rural Sociology (Chayanov, 1966) Agricultural History (Slicher van Bath, 1963) ↓ History of Technology (Giedion, 1969)
Industrial Economies and Societies	Economic History of Industrial Europe (Landes, 1969) Economic History of Industrial Japan (Rosovsky, 1966) Industrial Sociology (Bendix, 1956) Soviet and Other Communist Economies (Bergson, 1964) Economic Development of the Third World (Adelman and Morris, 1967) Modernization of the Third World (Mrydal, 1957)

money[15] and foreign trade.[16] In aboriginal stateless societies, a special province of anthropology (for example, precolonial New Guinea and North American Indians), one now also sees that external trade and a sort of social money (primitive valuables) were used in the formation of political and social alliances between clan segments.[17]

With regard to what economists call development and anthropologists and others call Third World modernization, enough theory and descriptive information exists now to answer questions such as how exactly were indigenous economies changed by European colonial rule; why did so little economic development occur under colonial rule; and how does economic development induce social change.[18]

There are also lessons to be learned about what an older generation of economists would have bluntly called methodology, but which Kuhn's *The Structure of Scientific Revolutions*, Popper's *The Open Society and Its Enemies*, and Rawls' *A Theory of Justice* are teaching us to call the philosophy of social science, including an answer to the question: what determines the degree of receptivity of each social science to Marxian theory? The answer, surely, is the quality of the explanatory power of whatever theory prevailed before Marxists began to challenge it. Marxism flourishes only where (a) no strong (that is, widely shared) earlier paradigm had been established, and (b) where quantitative evidence is difficult or impossible to adduce to demonstrate the explanatory power of analytical concepts and of analytical conclusions about real world structure and performance. Compared to anthropology and the other social sciences, economics has been much less affected by Marxian theory for these two reasons.

Finally, I will mention three research problems that I think economic anthropology can help with, but so far have not attracted much work by anthropologists.

The three worlds of economics, industrial capitalism, industrial communism, and the Third World of underdeveloped nations; *each* has star performers and poor performers for reasons not entirely explained by resource endowment, rates of investment and such, that is, for complicated historical, cultural, and institutional reasons.[19] Japan has been a star performer for one hundred years. Italy has not. Why? At least one Italian anthropologist has begun to work on such matters (see Carlo Tullio-Altan).

[15]See Grierson; P. Vidal-Naquet; Polanyi (1968a,b); the author (1965).

[16]See John K. Fairbank and S. Y. Teng; Fairbank (1942, 1968); Yi-T'ung Wang; Y. S. Yu; Percy Cooper Sands; Polanyi (1966, 1975).

[17]Polanyi (1957, 1968a); the author (1965, 1977a, 1978); Karl Heider.

[18]See Geertz; Epstein (1962, 1973); A. F. Holmberg; and the author (1978).

[19]See Myrdal (1957); Kurt Martin and John Knapp.

The second problem is best put as a question: What exactly is it in the structures of traditional tribal and peasant societies (in what are now independent countries of the Third World) that make socialist institutions and aspirations appealing to their intellectuals and political heads?

The last problem relates to rural development in the Third World. Anthropologists are grass roots experts who traditionally dwell in villages to carry out their professional research, and who provide us with detailed information about the structure of what exists in village communities. I think they have not been markedly effective so far in micro development—development from below, village development—because the traditional skills and theory of the anthropologist are seriously insufficient to address the complicated economic and technical problems of agricultural development and rural modernization today. But we do have a few good examples of economic anthropologists and teams of economists and anthropologists doing effective work on rural development.[20] Such work is only just beginning. Much much more remains to be done, particularly in theoretical formulation of how village development relates reciprocally to national development.

[20]See John Mellor et al.; Holmberg; Epstein (1962, 1973); De Wilde.

REFERENCES

G. Dalton, "Karl Polanyi's Analysis of Long-Distance Trade and his Wider Paradigm," in J. A. Sabloff and C. C. Lamberg-Karlovsky, eds., *Ancient Civilization and Trade*, Albuquerque 1975.

———, "Economic Anthropology," *Amer. Behav. Scientist*, May/June 1977, *20*, 635–56.

———, "The Impact of Colonization on Aboriginal Economies in Stateless Societies," in his *Research in Economic Anthropology*, Vol. I, Greenwich, CT 1978.

The Bazaar Economy: Information and Search in Peasant Marketing

By CLIFFORD GEERTZ*

There have been a number of points at which anthropology and economics have come to confront one another over the last several decades—development theory; preindustrial history; colonial domination. Here I want to discuss another where the interchange between the two disciplines may grow even more intimate; one where they may come actually to contribute to each other rather than, as has often been the case, skimming off the other's more generalized ideas and misapplying them. This is the study of peasant market systems, or what I will call bazaar economies.

There has been by now a long tradition of peasant market studies in anthropology. Much of it has been merely descriptive—inductivism gone berserk. That part which has had analytical interests has tended to divide itself into two approaches. Either the bazaar is seen as the nearest real world institution to the purely competitive market of neoclassical economics—"penny capitalism"; or it is regarded as an institution so embedded in its sociocultural context as to escape the reach of modern economic analysis altogether. These contrasting approaches have formed the poles of an extended debate between economic anthropologists designated "formalists" and those designated "substantivists," a debate that has now rather staled for all but the most persevering.

Some recent developments in economic theory having to do with the role of information, communication, and knowledge in exchange processes (see Michael Spence; George Stigler; Kenneth Arrow; George Akerlof; Albert Rees) promise to mute this formalism-substantivism contrast. Not only do they provide us with an analytic framework more suitable to understanding how bazaars work than do models of pure competition; they also allow the incorporation of sociocultural factors into the body of discussion rather than relegating them to the status of boundary matters. In addition, their actual use on empirical cases outside the modern "developed" context may serve to demonstrate that they have more serious implications for standard economic theory and are less easily assimilable to received paradigms than at least some of their proponents might imagine. If this is so, then the interaction of anthropology and economics may come for once to be more than an exchange of exotic facts for parochial concepts and develop into a reciprocally seditious endeavor useful to both.

*The Institute for Advanced Study.

I

The bazaar economy upon which my discussion is based is that of a town and countryside region at the foot of the Middle Atlas in Morocco I have been studying since the mid-1960's. (During the 1950's, I studied similar economies in Indonesia. See the author, 1963.) Walled, ethnically heterogeneous, and quite traditional, the town is called Sefrou, as is the region, and it has been there for a millenium. Once an important caravan stop on the route south from Fez to the Sahara, it has been, for about a century, a thriving market center of 15,000–30,000 people.

There are two sorts of bazaar there: 1) a permanent one, consisting of the trading quarters of the old town; 2) a periodic one, which meets at various spots—here for rugs, there for grain—outside the walls on Thursdays, as part of a very complex regional cycle involving various other market places and the other days of the week. The two sorts of bazaar are distinct but their boundaries are quite permeable, so that in-

dividuals move freely between them, and they operate on broadly the same principles. The empirical situation is extremely complex—there are more than 600 shops representing about forty distinct commercial trades and nearly 300 workshops representing about thirty crafts—and on Thursdays the town population probably doubles. That the bazaar is an important local institution is beyond doubt: two-thirds of the town's labor force is employed there.

Empirical detail aside (a full-scale study by the author is in press), the bazaar is more than another demonstration of the truth that, under whatever skies, men prefer to buy cheap and sell dear. It is a distinctive system of social relationships centering around the production and consumption of goods and services—that is, a particular kind of economy, and it deserves analysis as such. Like an "industrial economy" or a "primitive economy," from both of which it markedly differs, a "bazaar economy" manifests its general processes in particular forms, and in so doing reveals aspects of those processes which alter our conception of their nature. Bazaar, that Persian word of uncertain origin which has come to stand in English for the oriental market, becomes, like the word market itself, as much an analytic idea as the name of an institution, and the study of it, like that of the market, as much a theoretical as a descriptive enterprise.

II

Considered as a variety of economic system, the bazaar shows a number of distinctive characteristics. Its distinction lies less in the processes which operate and more in the way those processes are shaped into a coherent form. The usual maxims apply here as elsewhere: sellers seek maximum profit, consumers maximum utility; price relates supply and demand; factor proportions reflect factor costs. However, the principles governing the organization of commercial life are less derivative from such truisms than one might imagine from reading standard economic textbooks, where the passage from axioms to actualities tends to be rather nonchalantly traversed. It is those principles—matters less of utility balances than of information flows—that give the bazaar its particular character and general interest.

To start with a dictum: in the bazaar information is poor, scarce, maldistributed, inefficiently communicated, and intensely valued. Neither the rich concreteness or reliable knowledge that the ritualized character of nonmarket economies makes possible, nor the elaborate mechanisms for information generation and transfer upon which industrial ones depend, are found in the bazaar: neither ceremonial distribution nor advertising; neither prescribed exchange partners nor product standardization. The level of ignorance about everything from product quality and going prices to market possibilities and production costs is very high, and much of the way in which the bazaar functions can be interpreted as an attempt to reduce such ignorance for someone, increase it for someone, or defend someone against it.

III

These ignorances mentioned above are *known* ignorances, not simply matters concerning which information is lacking. Bazaar participants realize the difficulty in knowing if a cow is sound or its price right, and they realize also that it is impossible to prosper without knowing. The search for information one lacks and the protection of information one has is the name of the game. Capital, skill, and industriousness play, along with luck and privilege, as important a role in the bazaar as they do in any economic system. They do so less by increasing efficiency or improving products than by securing for their possessor an advantaged place in an enormously complicated, poorly articulated, and extremely noisy communication network.

The institutional peculiarities of the bazaar thus seem less like mere accidents of custom and more like connected elements of a system. An extreme division of labor and localization of markets, heterogeneity of products and intensive

price bargaining, fractionalization of transactions and stable clientship ties between buyers and sellers, itinerant trading and extensive traditionalization of occupation in ascriptive terms—these things do not just co-occur, they imply one another.

The search for information—laborious, uncertain, complex, and irregular—is the central experience of life in the bazaar. Every aspect of the bazaar economy reflects the fact that the primary problem facing its participants (that is, "bazaaris") is not balancing options but finding out what they are.

IV

Information search, thus, is the really advanced art in the bazaar, a matter upon which everything turns. The main energies of the bazaari are directed toward combing the bazaar for usable signs, clues as to how particular matters at the immediate moment specifically stand. The matters explored may comprise everything from the industriousness of a prospective coworker to the supply situation in agricultural products. But the most persistent concerns are with price and quality of goods. The centrality of exchange skills (rather than production or managerial ones) puts a tremendous emphasis on knowing what particular things are actually selling for and what sorts of things they precisely are.

The elements of bazaar institutional structure can be seen in terms of the degree to which they either render search a difficult and costly enterprise, or facilitate it and bring its costs within practical limits. Not that all those elements line up neatly on one or another side of the ledger. The bulk have effects in both directions, for bazaaris are as interested in making search fruitless for others as they are in making it effectual for themselves. The desire to know what is really occurring is matched with the desire to deal with people who don't but imagine that they do. The structures enabling search and those casting obstructions in its path are thoroughly intertwined.

Let me turn, then, to the two most important search procedures as such: clientelization and bargaining.

V

Clientelization is the tendency, marked in Sefrou, for repetitive purchasers of particular goods and services to establish continuing relationships with particular purveyors of them, rather than search widely through the market at each occasion of need. The apparent Brownian motion of randomly colliding bazaaris conceals a resilient pattern of informal personal connections. Whether or not "buyers and sellers, blindfolded by a lack of knowledge simply grop[ing] about until they bump into one another" (S. Cohen, quoted in Rees, p. 110), is, as has been proposed, a reasonable description of modern labor markets, it certainly is not of the bazaar. Its buyers and sellers, moving along the grooved channels clientelization lays down, find their way again and again to the same adversaries.

"Adversaries" is the word, for clientship relations are not dependency relations, but competitive ones. Clientship is symmetrical, egalitarian, and oppositional. There are no "patrons" in the master and man sense here. Whatever the relative power, wealth, knowledge, skill, or status of the participants—and it can be markedly uneven—clientship is a reciprocal matter, and the butcher or wool seller is tied to his regular customer in the same terms as he to them. By partitioning the bazaar crowd into those who are genuine candidates for his attention and those who are merely theoretically such, clientelization reduces search to manageable proportions and transforms a diffuse mob into a stable collection of familiar antagonists. The use of repetitive exchange between acquainted partners to limit the costs of search is a practical consequence of the overall institutional structure of the bazaar and an element within that structure.

First, there is a high degree of spatial localization and "ethnic" specialization of trade in the bazaar which simplifies the process of finding clients considerably and

stabilizes its achievements. If one wants a kaftan or a mule pack made, one knows where, how, and for what sort of person to look. And, since individuals do not move easily from one line of work or one place to another, once you have found a particular bazaari in whom you have faith and who has faith in you, he is going to be there for awhile. One is not constantly faced with the necessity to seek out new clients. Search is made accumulative.

Second, clientelization itself lends form to the bazaar for it further partitions it, and does so in directly informational terms, dividing it into overlapping subpopulations within which more rational estimates of the quality of information, and thus of the appropriate amount and type of search, can be made. Bazaaris are not projected, as for example tourists are, into foreign settings where everything from the degree of price dispersion and the provenance of goods to the stature of participants and the etiquette of contact are unknown. They operate in settings where they are very much at home.

Clientalization represents an actor-level attempt to counteract, and profit from, the system-level deficiencies of the bazaar as a communication network—its structural intricacy and irregularity, the absence of certain sorts of signaling systems and the undeveloped state of others, and the imprecision, scattering, and uneven distribution of knowledge concerning economic matters of fact—by improving the richness and reliability of information carried over elementary links within it.

VI

The rationality of this effort, rendering the clientship relation dependable as a communication channel while its functional context remains unimproved, rests in turn on the presence within that relation of the sort of effective mechanism for information transfer that seems so lacking elsewhere. And as that relation is adversary, so is the mechanism: multidimensional intensive bargaining. The central paradox of bazaar exchange is that advantage stems from surrounding oneself with relatively superior communication links, links themselves forged in sharply antagonistic interaction in which information imbalances are the driving force and their exploitation the end.

Bazaar bargaining is an understudied topic (but see Ralph Cassady), a fact to which the undeveloped state of bargaining theory in economics contributes. Here I touch briefly on two points: the multidimensionality of such bargaining and its intensive nature.

First, multidimensionality: Though price setting is the most conspicuous aspect of bargaining, the bargaining spirit penetrates the whole of the confrontation. Quantity and/or quality may be manipulated while money price is held constant, credit arrangements can be adjusted, bulking or bulk breaking may conceal adjustments, and so on, to an astonishing range and level of detail. In a system where little is packaged or regulated, and everything is approximative, the possibilities for bargaining along non-monetary dimensions are enormous.

Second, intensiveness: I use "intensive" in the way introduced by Rees, where it signifies the exploration in depth of an offer already received, a search along the intensive margin, as contrasted to seeking additional offers, a search along the extensive. Rees describes the used car market as one in which intensive search is prominent as a result of the high heterogeneity of products (cars driven by little old ladies vs. taxicabs, etc.) as against the new car market, where products are considered homogeneous, and extensive search (getting new quotations from other dealers) predominates.

The prominence of intensive bargaining in the bazaar is thus a measure of the degree to which it is more like a used car market than a new car one: one in which the important information problems have to do with determining the realities of the particular case rather than the general distribution of comparable cases. Further, it is an expression of the fact that such a market rewards a "clinical" form of search (one which focuses on the diverging interests of

concrete economic actors) more than it does a "survey" form (one which focuses on the general interplay of functionally defined economic categories). Search is primarily intensive because the sort of information one needs most cannot be acquired by asking a handful of index questions of a large number of people, but only by asking a large number of diagnostic questions of a handful of people. It is this kind of questioning, exploring nuances rather than canvassing populations, that bazaar bargaining represents.

This is not to say that extensive search plays no role in the bazaar; merely that it is ancillary to intensive. Sefrou bazaaris make a terminological distinction between bargaining to test the waters and bargaining to conclude an exchange, and tend to conduct the two in different places: the first with people with whom they have weak clientship ties, the second with people with whom they have firm ones. Extensive search tends to be desultory and to be considered an activity not worth large investments of time. (Fred Khuri reports that in the Rabat bazaar, bazaaris with shops located at the edge of the bazaar complain that such shops are "rich in bargaining but poor in selling," i.e. people survey as they pass, but do their real bargaining elsewhere.) From the point of view of search, the productive type of bargaining is that of the firmly clientelized buyer and seller exploring the dimensions of a particular, likely to be consummated transaction. Here, as elsewhere in the bazaar, everything rests finally on a personal confrontation between intimate antagonists.

The whole structure of bargaining is determined by this fact: that it is a communication channel evolved to serve the needs of men at once coupled and opposed. The rules governing it are a response to a situation in which two persons on opposite sides of some exchange possibility are struggling both to make that possibility actual and to gain a slight advantage within it. Most bazaar "price negotiation" takes place to the right of the decimal point. But it is no less keen for that.

REFERENCES

G. A. Akerlof, "The Market for 'Lemons': Quality, Uncertainty and the Market Mechanism," *Quart. J. Econ.*, Aug. 1970, *84*, 488-500.

Kenneth J. Arrow, *The Limits of Organization*, New York 1974.

R. Cassady, Jr., "Negotiated Price Making in Mexican Traditional Markets," *Amer. Indigena*, 1968, *38*, 51-79.

Clifford Geertz, *Peddlers and Princes*, Chicago 1963.

———, "Suq: The Bazaar Economy in Sefrou," in Lawrence Rosen et al., eds., *Meaning and Order in Contemporary Morocco: Three Essays in Cultural Analysis*, New York forthcoming.

F. Khuri, "The Etiquette of Bargaining in the Middle East," *Amer. Anthropologist*, July 1968, *70*, 698-706.

A. Rees, "Information Networks in Labor Markets," in David M. Lamberton, ed., *Economics of Information and Knowledge*, Hammondsworth 1971, 109-18.

M. Spence, "Time and Communication in Economic and Social Interaction," *Quart. J. Econ.*, Nov. 1973, *87*, 651-60.

G. Stigler, "The Economics of Information," in David M. Lamberton, ed., *Economics of Information and Knowledge*, Hammondsworth 1971, 61-82.

Towards a Marriage Between Economics and Anthropology and a General Theory of Marriage

By AMYRA GROSSBARD*

Historically, there have been clear lines of demarcation between economics and anthropology. Mary Douglas, p. 781, asserts that centripetal forces attract resources towards the center of a discipline and discourage turbulence at the boundaries of a subject out of fear of losing autonomy. If she is correct, then the present division of the social sciences may not be more than a historical accident, another instance of institutional self-perpetuation.

This paper is a declaration of turbulence. Building on the present trend to stretch disciplinary boundaries, it proposes a unification between economics and anthropology. As a first step, it is suggested that our disciplines could jointly work towards a general study of marriage.

I

Economics has traditionally explored the more quantifiable sectors of society with increasingly sophisticated theoretical and empirical tools. If we conceive social reality as a series of fields, economists generally farmed the scientifically most reachable ones at the intensive margin. Anthropologists, on the other hand, worked at the extensive margin of social science. Attempting to study entire cultures and venturing into the most remote communities, they have accumulated comprehensive insights at the expense of scientific methodology.

Recently, both economics and anthropology are extending their traditional boundaries: anthropology has become more concerned with the quantitative intensive margin, while economists have become more interested in the qualitative extensive margin of inquiry. Since ethnographies have been collected on most existing cultures, anthropologists have become more involved in cultural comparisons and theoretical generalizations. Anthropologist Ronald Cohen, who introduces a major methodological handbook by remarking that "the discipline as a whole does not have a systematic and cumulative tradition of methodological endeavor" (1973, p. v), expresses a "desire to see anthropology become a progressively more rigorous and scientific branch of the social sciences" (1973, p. vi), "our primary goal . . . is theory-construction" (1973, p. viii). In the same volume he also proposes "a restructuring of the social sciences [which] calls for methodological openness and a lack of concern for disciplinary boundaries" (1973, p. 49); Douglas specifically proposes that "economic analysis . . . be established at the centre of anthropology itself" (p. 782). She sees "the need for a cost-benefit analysis that would apply across the board to both monetary and non-monetary transactions" (p.781).

While more and more anthropologists concentrate less upon field work and more on theory and methods of analysis, economic investigation has expanded into the traditional specialities of other social sciences. For example, economic research has penetrated into the domain of the family (see for instance Theodore Schultz) and social interactions (see Gary Becker, 1976) and is even creating a link with sociobiology (see Jack Hirshleifer; Becker, 1976).

Are these contemporary developments in the two disciplines related? Are these

*Assistant professor, Occidental College. Thanks are given to Laura Bogden, Ronald Cohen, John Edwards, Joel Guttman, Jonathan Leland, Scott Littleton, and Theodore W. Schultz for helpful comments.

hands reaching out to any encounter? It seems that these new trends presage the emergence of a social science unified in its efforts to understand man, society, and culture. Marriage can serve as a good illustration of what social science stands to gain if economics and anthropology join forces.

II

When economics overcomes its preoccupation with monetary dimensions, it can deal with any predominantly nonmonetary transaction, including marriage. The economic analysis of marriage is based on working assumptions that have proved convenient in the past but can be rejected if they ever lose their explanatory power. These assumptions view people first as rational maximizers of utility, and second as substitutable. In its application to marriage, this means that individuals make optimal choices regarding whom and when they marry, and that, because substitution between potential spouses is possible, interdependence of individual decision making takes the form of a market mechanism.[1]

In making implicit or explicit cost-benefit analysis of marriage and divorce, individuals are guided by utility functions that depend on both culture and nature. The nonmonetary essence of the gains from marriage impedes measurement and leads the economist to focus on separate determinants of such gains: for instance (see Becker, 1973) the complementarity betwen spouses (i.e., in producing own children); the relative productivity of men and women inside and outside the home; income; age; education and other traits affecting productivity; and demand for a particular composition of marital output. The first empirical studies of marriage by economists focused on the contemporary United States, looking at the causes for differences in percentage of married women per state, individual age at marriage and probability of divorce. One of the findings that is perhaps counterintuitive but still consistent with economic theory is the inverse relation between the percentage of women married and of Catholic population across *U.S.* states. Alan Freiden's explanation relies on the expected costs of divorce: Catholic marriages are less profitable because of higher expected costs of divorce. A second finding was the larger gains from marriage among individuals with higher income: *ceteris paribus* they marry earlier. (General impressions about earlier marriage among the poor derive from inaccurate statistical inference.) Becker, Elizabeth Landes and Robert Michael add another piece of evidence for the positive income effect on marriage: American marriages are less likely to dissolve when income is higher, although wealth exceeding the level expected at time of marriage may also increase the chances of dissolution. This follows since divorce is dependent on uncertainty as well as gain from marriage.

Such studies contribute to a social science of marriage because of the simplicity of theory and the sophistication of econometric methods. They circumvent the question of content of the utility function by studying differences among poeple who live in the same culture and who have probably adopted the same values. However, a social science of marriage needs to know more about the meaning of marital behavior, that is, explore the content of the utility functions. The above-mentioned study of divorce also illustrates the effect of a clearly cultural factor, religion, on marital dissolution. In making cross-cultural generalizations, especially when cultures are far apart, we need a deeper understanding of marriage, which has been precisely the preoccupation of anthropologists.

III

In addition to gathering huge numbers of facts about marriage around the world, anthropologists have dealt theoretically with this issue. Two major schools of anthropologists have discussed determinants of utility from marriage. Rather

[1] The assumption of substitutability may be untenable in a society with prescribed marriages.

like sociobiologists, *functionalists* view marriage as a means to satisfy functions like reproduction, socialization, and transmission. Some stress the function of marriage in meeting the needs of other parts of the social system while other emphasize physical needs like sexual gratification. This predominantly British approach, which leads to institutional determinism and encourages ethnocentrism, reached its peak of popularity before England lost its colonial empire. *Structuralists* disagree with the emphasis on nature and society as determinants of marriage; they think that cultural factors like relative reliance on the capacity to reason generate variations in the meaning (utility) individuals attribute to identical activities. Their analysis draws increasingly on linguistics, since language can be considered as a major expression of collective meaning (see James Boon and David Schneider). Besides these two major schools, there have been evolutionary theories, ecological analyses illustrating the importance of the physical environment on the structure of marriage and descent, and Marxist analyses stressing the important effect of means of production. While the theoretical focus of functionalists and structuralists centers on meaning and utility, the latter two approaches point out constraints in the real world affecting individual and community choice—a view compatible with economic theory. Structuralism may also share an assumption with economics: Claude Levi-Strauss' binary oppositions built in the structure of the human mind and creating universal components of culture appear consistent with the economist's concept of cost vs. benefit.

Economics thus provides an umbrella theory of marriage that creates common ground within the anthropology of marriage. For example, the content of utility functions often does not matter when you compare culturally homogeneous units—an important message from economics that could help integrate fascinating ethnographic material. If differences of opinion between anthropologists become more systematic, one could accept parts of an anthropologist's empirical findings and generalizations, and simultaneously disagree about other parts of the analysis. Jointly, anthropologists and economists could 1) focus their talents on the most difficult questions (utility for instance) taking advantage of their respective skills; 2) give new significance to previous ethnographic findings; and 3) collect better data. The last two points are illustrated in the next section.

IV

To illustrate how a general theory of marriage can reinterpret evidence collected by anthropologists, let me use examples out of my own work on polygyny. In applying regression analysis to Cohen's own data, I found that women at peak fecundity have fewer cowives, which follows from a substitution between quantity and quality of wives, a point that had not occurred to Cohen before I drew attention to it (see the author, p. 103).

The legal imposition of monogamy can be viewed as an interference in the marriage market curtailing the aggregate demand men have for wife services, whatever that means within the specific culture. Consequently, assuming the supply of wife services by women has not changed, the new equilibrium (mainly nonmonetary) wife wage will be less advantageous to women. Societies that impose monogamy therefore harm the welfare of women by reducing male competition for their services. Three kinds of evidence show that women are better off when polygyny is permitted: use of brideprice, age at marriage, and proportion of women married.

Dowry and brideprice (a transfer from the husband to the bride's family) reflect monetary dimensions of the wife wage at time of marriage. The presence of a dowry (negative brideprice) system probably reflects a lower wife wage, for it means that a woman (or her family) has to pay for the privilege of getting married. Cross culturally, dowry is strongly linked with monogamous (and polyandrous) marriage while the institution of brideprice (a wealth

transfer from the husband to the bride's family) is more often found in polygynous than in monogamous African societies. Inspired by a trip to India, Martin Bronfenbrenner has also written that a positive brideprice is more likely when the number of wives per husband exceeds unity.

Not only does it seem that the probability of finding brideprice vs. dowry varies directly with the presence of polygyny, but evidence also suggests that brideprice payments are higher in more polygynous societies. Comparing two Sebei communities in eastern Uganda, anthropologist Walter Goldschmidt found that the brideprice was considerably higher in the more polygynous community, p. 316. Encouraged by his findings among the Sebei, Goldschmidt then examined thirteen separate societies in East Africa and obtained a clear simple correlation between polygyny and brideprice levels, p. 327.

When women gain more from marriage, they are also likely to marry younger. This seems indeed to be the case: women's average age at marriage is 13 or 14 among the Hausa and the Kanuri of eastern Nigeria, societies with widespread polygyny. The Tallensi, another West Africa tribe, are slightly less polygynous and their daughters marry somewhat later. Here, the average female age at marriage is 16 and 17. On the whole, women marry considerably earlier in polygynous areas like Africa and the Moslem world than in monogamous Europe and America. Polygyny also raises the difference in mean age at marriage of men and women. Using data from sixteen districts of the Congo, William Brass et al. found a simple correlation of .8 between an index of polygyny and difference in mean age at marriage.

While in the United States the husband is on average two years older than the wife, that difference rises to seven years in the Arab world and to ten years in some heavily polygynous African societies like the Kanuri.

Not only will polygyny encourage women to marry earlier, but it will lead a larger proportion of women to marry at all ages. Comparing two ethnic groups in the Ivory Coast Remi Clignet found more unmarried females in the less polygynous tribe, p. 110. Likewise, better marital income opportunities open to women will lead widows to accept being "inherited" by relatives of their husbands, as is specified in the institution of "levirate." Using the same data from the Congo, Brass et al. showed a negative correlation of −.45 between polygyny and the proportion of widowed and divorced among women 15 to 45 years old. Although none of those separate facts necessarily demonstrates the benefits of polygyny to women, such benefits become more real in light of quantity and variety of evidence.

A general theory not only associates previously unrelated facts, like brideprice and age at marriage, it also points out variables on which data should be gathered. For instance, in my attempt to explain the number of wives present in Maiduguri households, I found male Koranic education to increase the number of a man's wives, while the same education obtained by females reduced the number of co-wives. This statistical finding, based on data other than Cohen's concorded with an economic theory of polygyny and led Cohen to regret not having included religious schooling in his own questionnaire.[2]

The present state of the general theory of marriage is definitely unsatisfactory. Most economists who promote it have been limited to the American experience and have not sufficiently questioned the rationale behind institutional constraints. Polygyny is only one example. What leads to the existence of institutions like the levirate, patrilineality, and dowry? What effects do they have on fertility, labor force participation, divorce, or other variables with important policy implications? Without massive help from kinship specialists, a really general theory of marriage cannot be established.

V

The creation of a common language and method is necessary to extend the intensive

[2]Related in a personal communication to the author.

and intensify the extensive, thus building a science that combines the robustness of theories and empirical work with broad cultural perspectives. While most social sciences will eventually become involved in that unification process, it does not seem coincidental that explicit or implicit promoters of such a marriage come from our two disciplines in particular: the gains from marriage are larger the more each partner complements the other. All disciplines can be viewed as potential partners in a marriage market. If it is true that a combination of extensive and intensive perspectives enriches social science, disciplines with the largest variation in intensive vs. extensive productivity have the most to gain from marriage. Since anthropology and economics lie respectively at the extensive and intensive ends of the spectrum, their gains from marriage will be particularly high.

This unified view on marriage represents only one possible direction of such interscience marriage. It is a good starting point, not only for its symbolism, but also because cooperation between economics and anthropology has long been hindered by lack of applicability of economics to small scale traditional societies or perception of such lack of applicability. Economics is now changing by involving itself with marriage and other more human and less monetary transactions. This new brand of "human economics" can contribute to the analysis of all human behavior that potentially involves use of reason and rational decision making. Economics aspires to be a "science of man"—English for "anthropology."

REFERENCES

G. S. Becker, "A Theory of Marriage," *J. Polit. Econ.*, July 1973, *81*, 813–46.

——, *The Economic Approach to Human Behavior*, Chicago 1976.

——, **E. Landes, and R. Michael.** "An Economic Analysis of Marital Instability," *J. Polit. Econ.*, Dec. 1977, *85*, 1141–88.

J. A. Boon and D. M. Schneider, "Kinship Vis-à-Vis Myth: Contrasts in Levi-Strauss' Approachs to Cross-Cultural Comparison," *Amer. Anthropologist*, Dec. 1974, *76*, 799–817.

William Brass et al., *The Demography of Tropical Africa*, Princeton 1968.

M. Bronfenbrenner, "A Note on the Economics of the Marriage Market," *J. Polit. Econ.*, Nov./Dec. 1971, *79*, 1424–25.

Remi Clignet, *Many Wives, Many Powers*, Evanston 1970.

R. Cohen, "Preface" and "Generalizations in Ethnology," in Raoul Naroll and Ronald Cohen, eds., *A Handbook of Method in Cultural Anthropology*, New York 1973.

——, "On Grossbard's Economic Analysis of Polygyny in Maiduguri," *Curr. Anthropology*, Mar. 1977, 102–05.

M. Douglas, "The Exclusion of Economics," *Times Literary Suppl.*, July 6, 1973, 781–82.

A. Freiden, "The U.S. Marriage Market," in Theodore W. Schultz, ed., *Economics of the Family*, Chicago 1974.

W. Goldschmidt, "The Economics of Brideprice among the Sebei and in East Africa," *Ethnology*, Oct. 1974, *13*, 311–31.

A. Grossbard, "An Economic Analysis of Polygyny: The Case of Maiduguri," *Curr. Anthropology*, Dec. 1976, *17*, 701–07.

J. Hirshleifer, "Competition, Cooperation and Conflict in Economics and Biology," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 238–43.

Claude Lévi-Strauss, *The Raw and the Cooked: Introduction to a Science of Mythology*, Vol. 1, New York 1969.

Theodore W. Schultz, *Economics of the Family*, Chicago 1974.

Unemployment in Capitalist Regulated Market Economies and Socialist Centrally Planned Economies

By MORRIS BORNSTEIN*

The comparative economic systems literature commonly distinguishes two major types of economic systems—capitalist regulated market economies (*CRME*) and socialist centrally planned economies (*SCPE*).

The *CRME* have several key characteristics: 1) most of the means of production are privately owned; 2) market forces chiefly determine the level of economic activity, the rate of growth, the composition of output, and the distribution of income. But 3) the government intervenes in the economy in various ways to deal with problems of growth, monopoly, inflation, unemployment, income distribution, etc. The *CRME* include, for example, the United States, Canada, Western Europe, Japan, and Australia.

In the *SPCE*, on the other hand, 1) most means of production are collectively owned, and 2) a large administrative bureaucracy attempts to direct resource allocation and income distribution through comprehensive and detailed planning. However, 3) the market mechanism is used to distribute the labor force among planned jobs, and to distribute the planned supply of consumer goods among households, which exercise consumer choice in the expenditures of their money incomes on available goods at prevailing prices. Among the *SCPE* are the *USSR*, most East European countries, and the People's Republic of China.

This paper compares the two systems in regard to types of unemployment and antiunemployment measures. The experience of particular countries such as the United States or the *USSR* is mentioned by way of illustration, but it is not suggested that they are prototypical or ideal examples of their respective economic systems. There are interesting differences in regard to unemployment among the *CRME* and among the SCPE, arising from differences in the level of economic development, the structure of the economy, the demographic composition of the labor force, and social and institutional arrangements, but space limitations preclude discussion of such intragroup differences (for the *CRME*, see, for instance, Joyanna Moy and Constance Sorrentino). Also, lack of relevant data for the *SCPE* prevents reliable statistical comparisons of unemployment rates in the two systems (for a valiant attempt, see P.J.D. Wiles). Finally, the paper does not consider unemployment in a third type of economic system—a socialist regulated market economy—because the sample of countries (Yugoslavia) is too small to permit sound generalizations about characteristics of such a system which would transcend particular national circumstances.

I. Types of Unemployment

Unemployment may be analyzed in terms of its origin or its form. The first ap-

*Professor of economics, University of Michigan. This paper draws on a research project at the University of Michigan Center for Russian and East European Studies funded by a grant from the Rockefeller Foundation. I wish to thank the Foundation for its support; Dennis A. O'Hearn for his assistance in research; and Robert F. Dernberger, Philip Hanson, and Harold M. Levinson for valuable comments.

proach is more common but the second is perhaps more illuminating in comparing economic systems.

A. *Origin*

In the *CRME* four kinds of unemployment are commonly identified. *Seasonal* unemployment arises from regular seasonal fluctuations in the demand for and supply of labor. On the demand side, seasonal variations in labor requirements are common, for instance, in agriculture, construction, and vacation resort activities. On the supply side, the school calendar affects the labor force participation of young people. *Frictional* unemployment involves those seeking their first jobs and those between jobs after quitting or being dismissed. *Structural* unemployment occurs when people are unable over a long period to find work, because of a lack of skills or a shortage of jobs in their particular labor market areas, although national aggregate demand may be high. Finally, *cyclical* unemployment results from a general decline in business activity, with a subsequent recovery leading to reemployment.

In the *CRME*, seasonal and frictional unemployment are considered part of the normal process by which the market mechanism adjusts production to demand. Thus a seasonally adjusted frictional unemployment rate of, say, 3–4 percent of the labor force could be considered compatible with "full employment." On the other hand, cyclical and structural unemployment have more serious economic and social consequences and more urgently require corrective government action.

The *SCPE* acknowledge seasonal unemployment as inevitable due to natural factors, such as agricultural production cycles, and institutional arrangements, like the academic calendar (see P.E. Strelets). In turn, some frictional unemployment is expected so long as new entrants into the labor force can choose their first jobs, workers are free to quit, or enterprise managers are permitted to dismiss unsatisfactory or surplus workers. But the authorities in the *SCPE* often complain about job changes on the initiative of workers—stressing the costs to the economy in lost production during periods of unemployment, in low productivity when workers begin new jobs, and in retraining programs (see L. Kuprienko).

In the *SCPE*, structural unemployment is regarded as evidence of deficiencies in planning, which is supposed to balance the supply of and demand for labor through accurate estimates of the labor force and correct decisions on training programs, technological progress, location of new capacity, and relative wages. This is an extremely difficult task (see Gertrude Schroeder), and in the *USSR*, for instance, structural unemployment has persisted in certain regions (see A.V. Topilin, p. 86) and among women in smaller cities and rural areas (see O. Latifi).

In contrast, the *SCPE* do not show the recurrent pattern of cyclical unemployment common in the *CRME*. The reasons include the *SCPE*s' "taut" planning for ambitious growth targets and their measures to insulate the domestic economy from external shocks (see Franklyn Holzman, 1968). Aside from harvest variations, fluctuations in economic activity in the *SCPE* result primarily from changes in growth rates of investment and output, without periodic recessions in which workers are laid off (see Alexander Bajt).

B. *Form*

The distinction between "open" and "disguised" unemployment in some ways parallels the more familiar contrast between open and "repressed" inflation. The openly unemployed are seeking jobs but cannot find them. Disguised unemployment—or underemployment—occurs when workers have jobs but are underutilized because they wish full-time jobs but can get only part-time work; because their full-time jobs do not use all their skills and training; or because, though employed full-time in jobs matching their qualifications, their productivity is low.

The first kind of disguised unemployment—part-time work when full-time jobs are sought—is more common in the *CRME* than in the *SCPE*. In the latter, part-time work is more often a way to draw into the labor force pensioners, housewives, and others not available for full-time jobs.

The second category—work below skill level—occurs in both systems because training opportunities and choices are not correctly matched with (future) employment possibilities, information on vacancies is imperfect, and mobility is limited.

The causes of low productivity may include late delivery of materials, equipment breakdowns, poor organization of production, low output norms, weak incentives due to the level and structure of compensation, and hoarding of labor by enterprise management. All of these may be found in both the *CRME* and the *SCPE*, but space limitations permit discussion only of labor hoarding, about which recent literature has shown some interesting similarities between the two systems.

In the *SCPE* such hoarding is explained by the combination of the "ratchet" principle in planning, tight labor market conditions, and restrictions on the dismissal of redundant workers. The first makes management expect that already ambitious production assignments will be "jacked up" continually in future periods. The second makes it believe that it may be difficult to replace workers who quit or to hire additional workers authorized for the extra output. Therefore, management wants a pool of "reserve" workers on the payroll, not because they are needed now but because they are likely to be required, but perhaps unavailable, in the future. Finally, there may be restrictions of law or custom on management's ability to discharge unneeded workers (see George Feiwel, pp. 346–54; Joseph Berliner, pp. 165–67; *Osnovnye problemy*, pp. 14–20).

Labor hoarding also exists in the *CRME* when labor input is not adjusted fully to changes in output. Inaccurate sales forecasts can cause "unplanned" hoarding. Conscious "planned" hoarding occurs because of legal commitments, indivisibilities in production, morale considerations, transaction costs in firing and hiring, and expectations of future changes in demand (see Stuart McKendrick, and K.G. Knight and R.A. Wilson). The Japanese version is sometimes called "permanent employment" (see Robert Cole).

Thus, a comparison of unemployment in *CRME* and *SCPE* must consider both the open and the disguised forms. One may hypothesize that central planning's effort to mobilize resources for rapid growth and socialism's ideological commitment to the "right to work" tend to hold open unemployment in *SCPE* below the levels observed in *CRME*. At the same time, these two factors may lead to more disguised unemployment in *SCPE* than in *CRME*.

Unfortunately, the statistical data to test such hypotheses are lacking. However, the relative importance of the two forms of unemployment may be inferred from the focus of policy discussions and antiunemployment measures in the two systems. In the *CRME*, the chief concern is with open unemployment, while the *SCPE* pay much more attention to underemployment. Thus, the *SCPE* stress that "full employment" means not only 1) that there is a job for everyone who wants one, but also 2) that labor should be allocated rationally across the economy, and 3) that it should be used efficiently inside the enterprise (see V.P. Korchagin, p. 35, and I. Ushkalov). Similarly, whereas the concept of "labor reserve" has been applied in the United States to refer to potential workers currently outside the labor force (see Christopher Gellner), in the *USSR* it encompasses as well "hidden" or "internal" reserves in the labor input of the employed but underutilized (see *Osnovnye problemy*, pp. 231–32).

II. Antiunemployment Measures

In turn, the two systems differ in their approach to reducing unemployment.

A. *The CRME*

In the *CRME* there are two types of constraints on antiunemployment measures. At the micro level, society is loath to impose restrictions on the freedom of workers to choose and quit jobs and on the freedom of firms to determine the amount and kind of labor input. At the macro level, it is widely believed that a reduction in unemployment will be accompanied by an increase in inflation. Thus, policymakers try to determine a "tolerable" rate and composition of unemployment. Efforts to reduce unemployment include job creation, matching the unemployed with job vacancies, and reducing the size of the "labor force" desiring employment.

In the *CRME* the conventional approach has been to create jobs in the private sector by stimulating aggregate demand through expansionary monetary policies, tax cuts, and increases in government purchases of goods and services. More recently, some *CRME* governments have subsidized part of private firms' wage costs to induce them to maintain the employment of workers who would otherwise be laid off, or to hire additional workers from the unemployed (see John Burton). Finally, additional (temporary or permanent) public sector jobs can be established at the national, regional, or local levels (see Michael Wiseman).

The matching of unemployed workers with existing or newly created job vacancies is attempted through placement services, retraining, relocation assistance, and area development schemes.

The unemployment rate can also be decreased by reducing the size of the labor force wishing paid employment. Labor force participation of the "old" can be cut by lowering the retirement age and by raising pensions so fewer retirees seek work. In turn, the entry of the "young" into the labor force can be delayed by extending compulsory secondary education and by increasing tuition and maintenance subsidies for postsecondary training. Finally, as in West Germany for example, the number of foreign "guest workers" can be cut to expand employment opportunities for citizens.

In addition, the effects of unemployment on household income may be at least partially offset by unemployment insurance schemes and welfare programs.

B. *The SCPE*

In the *SCPE* the attainment of the political leaders' goal of full mobilization of labor resources is constrained by the extent of workers' freedom to quit—much greater in Eastern Europe than in the People's Republic of China—as well as by the inability of planning agencies to control enterprise operations and regulate labor markets exactly as they wish. But the inflation-unemployment tradeoff is regarded with much less concern than in the *CRME*. The *SCPE* are more confident about their ability to curtail inflationary pressure by tax and credit measures or to repress it by comprehensive price controls.

In the *SCPE* jobs are created by ambitious development plans implemented by detailed administrative orders, government expenditure programs, and an accommodating monetary policy. The socialist character of the economy resolves the issue of public vs. private sector employment overwhelmingly in favor of the former—with the private sector usually limited to subsidiary agriculture, handicrafts, and selected personal services. Explicit wage subsidies to public enterprises are the exception rather than the rule. But firms attempt to cover the cost of surplus labor by negotiating larger wage bills in their annual plans and by securing higher prices from central agencies, which usually set prices to cover planned branch average cost plus a profit markup on it.

For matching workers with job vacancies, the *SCPE* emphasize training programs, and organized mass recruitment for new towns. Although employment bureaus are useful for the latter, the official attitude toward them in *SCPE* is ambivalent. The authorities fear that such

bureaus may not only help in centrally planned labor placement, but may also facilitate enterprises' efforts to build up "reserve" labor pools and "unplanned" labor turnover at the initiative of workers. Thus, in the *USSR* for instance, these bureaus are supposed to screen all requests for workers and reject those inconsistent with an enterprise's labor plan, and one indicator of the bureaus' performance is how long workers stay in the jobs in which they are placed (see E.V. Kasimovskii, pp. 122–26).

In the *SCPE* the relocation of workers to new areas is difficult. One factor is the social and cultural obstacles to mobility similar to those in the *CRME*. Housing shortages are widespread and persistent because of investment plans' neglect of infrastructure. Also, the high participation rate of married women in the labor force requires suitable jobs for both spouses in the new city. Thus, except where location decisions are determined by natural factors like mineral deposits, it has usually proved more feasible to place new capacity in labor surplus areas than to relocate workers.

The *SCPE* regard unemployment insurance as unnecessary and indeed harmful. A high level of aggregate demand is supposed to assure enough jobs for all. Furthermore, when a worker is no longer needed by a firm, because of changes in output assignments or production methods (for example, mechanization or automation), the management is expected, or even legally obligated, if possible to reassign him, with retraining if necessary, inside the enterprise; or to arrange for placement in another firm in the same area. Therefore, the authorities consider that unemployment is voluntary, rather than involuntary, and that compensation for it is not only unjustified but could extend the unemployment period by financing a longer job search. (A similar view may be found in the recent discussion in the United States about the disincentive effects of unemployment compensation, summarized in Gary Fields, and Arnold Katz and Joseph Hight.)

III. Conclusion

Contrary to some official claims, unemployment does occur in the *SCPE*—because of nature, workers' freedom of job choice, and imperfect planning. Although seasonal, frictional, and structural unemployment exist in the *SCPE*, they do not have cyclical unemployment problems comparable to those in the *CRME*. However, the distinction between open and disguised unemployment—alternative forms of inefficiency—is important. In the *SCPE*, a high level of aggregate demand due to ambitious national plans can create a labor shortage at the macro level, while there are labor surpluses at the micro level as a result of the firm's reaction to taut plans and tight labor markets.

These differences are in turn reflected in the emphasis of antiunemployment measures in the two systems. The *CRME* stress job creation, with controversy over the shares for the private and public sectors. The *CRME* maintain income and support the job search of the unemployed through unemployment compensation. In the *SCPE*, the authorities consider unemployment compensation unnecessary and detrimental, and they try through demanding plans to mobilize "internal" labor reserves of the employed but underutilized.

REFERENCES

A. Bajt, "Investment Cycles in European Socialist Economies: A Review Article," *J. Econ. Lit.*, Mar. 1971, *9*, 53–63.

Joseph S. Berliner, *The Innovation Decision in Soviet Industry*, Cambridge, Mass. 1976.

J. Burton, "Employment Subsidies—the Cases For and Against," *Nat. Westminster Bank Quart. Rev.*, Feb. 1977, 33–43.

R. E. Cole, "Permanent Employment in Japan: Facts and Fantasies," *Ind. Labor Relat. Rev.*, Oct. 1972, *26*, 615–30.

G. R. Feiwel, "Causes and Consequences of Disguised Industrial Unemployment

in a Socialist Economy," *Soviet Stud.*, July 1974, *26*, 344–62.

G. S. Fields, "Direct Labor Market Effects of Unemployment Insurance," *Ind. Relat.*, Feb. 1977, *16*, 1–14.

C. G. Gellner, "Enlarging the Concept of a Labor Reserve," *Mon. Labor Rev.*, Apr. 1975, *98*, 20–28.

F. D. Holzman, "Unemployment in Planned and Capitalist Economies: Comment," *Quart. J. Econ.*, Aug. 1955, *49*, 452–60.

——, "Soviet Central Planning and Its Impact on Foreign Trade Behavior and Adjustment Mechanisms," in Alan A. Brown and Egon Neuberger, eds., *International Trade and Central Planning: An Analysis of Economic Interactions*, Berkeley; Los Angeles 1968, 280–305.

E. V. Kasimovskii, *Trudovye resursy: formirovanie i ispol'zovanie* (*Labor Resources: Formation and Utilization*), Moscow 1975.

A. Katz and J. E. Hight, "The Economics of Unemployment Insurance: A Symposium —Overview," *Ind. Labor Relat. Rev.*, July 1977, *30*, 431–37.

K. G. Knight and R. A. Wilson, "Labor Hoarding, Employment, and Unemployment in British Manufacturing Industry," *Appl. Econ.*, Dec. 1974, *6*, 303–10.

V. P. Korchagin, *Trudovye resursy v usloviiakh nauchno-tekhnicheskoi revoliutsii* (*Labor Resources in Conditions of the Scientific-Technical Revolution*), Moscow 1974.

L. Kuprienko, "Vlianie urovnia zhizni na dvizhenie trudovykh resursov" ("The Influence of Living Standards on the Movement of Labor Resources"), *Voprosy ekonomiki*, 1972, no. 3, 22–31.

O. Latifi, *Pravda*, Apr. 20, 1977, p. 2; English translation in *Curr. Digest Soviet Press*, May 18, 1977, *29*, 24–25.

S. McKendrick, "An Inter-Industry Analysis of Labor Hoarding in Britain, 1953–72," *Appl. Econ.*, June 1975, *7*, 101–17.

J. Moy and C. Sorrentino, "An Analysis of Unemployment in Nine Industrial Countries," *Mon. Labor Rev.*, Apr. 1977, *100*, 12–24.

A. R. Oxenfeldt and E. van den Haag, "Unemployment in Planned and Capitalist Economies," *Quart. J. Econ.*, Feb. 1954, *48*, 43–60.

—— and ——, "Reply," *Quart. J. Econ.*, Aug. 1955, *49*, 461–64.

G. Schroeder, "Labor Planning in the USSR," *Southern Econ. J.*, July 1965, *32*, 63–72.

P. E. Strelets, *Problema sezonnosti truda v sotsialisticheskom sel'skom khoziaistve* (*The Problem of Seasonality of Labor in Socialist Agriculture*), Omsk 1973.

A. V. Topilin, *Territorial'noe pereraspredelenie trudovykh resursov v SSSR* (*Territorial Redistribution of Labor Resources in the USSR*), Moscow 1975.

I. Ushkalov, "Effektivnost' ispol'zovaniia trudovykh resursov v stranakh-chlenakh SEV (Obzor)" (Effectiveness of Utilization of Labor Resources in CMEA Countries (A Survey)), *Voprosy ekonomiki*, 1977, No. 4, 123–31.

P. J. D. Wiles, "A Note on Soviet Unemployment on U.S. Definitions," *Soviet Stud.*, Apr. 1972, *23*, 619–28.

M. Wiseman, "Public Employment as Fiscal Policy," *Brookings Papers*, Washington 1976, *1*, 67–104.

Osnovnye problemy ratsional'nogo ispol'zovaniia trudovykh resursov v SSSR (*Basic Problems of Rational Utilization of Labor Resources in the USSR*), Moscow 1971.

Unemployment in Western Europe and the United States: A Problem of Demand, Structure, or Measurement?

By ROBERT H. HAVEMAN*

High measured unemployment, often accompanied by rapid wage and price increases, has plagued most Western nations in the 1970's. The aim of this paper is to appraise, albeit crudely, the contribution of three factors to these patterns in the Netherlands, Sweden, the United Kingdom, and the United States. These factors are: 1) insufficient aggregate demand, 2) structural imbalances in the composition of labor supplies and demands, and 3) changes in the relationship of *measured* unemployment to excess labor supply (referred to as the *U-ES* relationship). My main thesis is that measured unemployment bears a different relationship to real excess labor supply in the 1970's than it did in the 1960's, explaining much of the increase in measured unemployment from the 1960's to the 1970's.

I. Some Unemployment, Wage, and Price Facts

Table 1 shows measured unemployment, 1952–76, for the four countries. Except for Sweden, unemployment rates in the 1970's are substantially higher than those of earlier periods. In addition, the nature of the unemployment problem has also changed. In both the United States and Western Europe, the long-duration unemployment rate has been secularly increasing. The *U.S.* rate rose from an average of .3 percent in the 1965–70 period to .6 percent in the 1970–75 period; for comparable periods, the *U.K.* change was from .5 to .8 percent. Except for Sweden, as recessions have turned to recovery, peak unemployment has crept up and estimated relationships between measured unemployment and the ratio of actual to full employment *GNP* have secularly increased.

The composition of unemployment has also changed in the 1970's, with the radical increase in youth unemployment being the most noteworthy. From 1970–76, the youth unemployment rate increased from 1.4 to 9.1 percent in the Netherlands, from 2.8 to 3.6 in Sweden, from 2.7 to 11.1 in the United Kingdom, and from 9.9 to 14.1 in the United States. The level of youth unemployment in the United States, with the most rapidly growing youth labor supply, stands significantly higher than that of the Western European countries.

Wage rate increases in the 1970's are also different from the earlier ones. For the European nations, post-1970 annual increases in excess of 15 percent are by no means rare, while 10 percent rates were regarded as exceptional in the 1960's. The post-1970's average increase is the greatest for the United Kingdom (16.6 percent), with the Netherlands (15.2 percent) a close second. By comparison, the average increase for the United States (7.2 percent) appears modest. Price increases parallel wage increases, with the United Kingdom (12.8 percent) experiencing the greatest post-1970's rate among the four countries.

*Professor of economics, University of Wisconsin-Madison, and fellow, Institute for Research on Poverty. The assistance of Patricia Capdevielle and Joyanna Moy of the U.S. Department of Labor and Terence Kelly of the Organization for Economic Cooperation and Development (*OECD*) in obtaining data on the countries studied is gratefully acknowledged. Most of the comparative statistics on the trend and composition of unemployment are from *OECD*. I would also like to thank Sheldon Danziger, Robert Flanagan, Irwin Garfinkel, Donald Hester, Rudolf Meidner, David Richardson, Eugene Smolensky, and Jacques van Der Gaag for helpful comments.

TABLE 1—UNEMPLOYMENT RATE PATTERNS FOR THE NETHERLANDS, SWEDEN, UNITED KINGDOM, AND THE UNITED STATES, 1952–76[a]

	Netherlands	Sweden	United Kingdom	United States
1952–59	2.02	2.06	1.75	4.51
1960–64	.86	1.58	1.97	5.76
1965–69	1.39	1.82	2.11	3.91
1970	1.18	1.53	2.62	4.43
1971	1.44	2.55	3.54	5.90
1972	2.40	2.78	3.97	5.75
1973	2.44	2.58	2.59	5.00
1974	2.96	1.95	2.59	5.13
1975	4.25	1.68	4.09	8.03
1976	4.63	1.60	5.65	7.78

[a]The data shown are averages of the quarterly aggregate unemployment rate for the period indicated. The data sources are described in an appendix, which is available from the author.

II. Demand, Structure, and Measurement: Some Speculations

These data are consistent with several conjectures involving shortages of aggregate demand, growing structural imbalances, increasing wage and price determination roles for rational expectations, exogenous shocks, and trade union power, and a changing relationship of measured unemployment to excess labor supply. The following focusses largely on the last of these.

In the absence of substantial changes in the composition of the supply of and demand for labor, or in constraints on the operation of the labor market, any increase in labor supply can be absorbed by an increase in aggregate demand. The greater the rate of labor force growth, of course, the greater the potential difficulty in maintaining full employment. While labor force growth rates in the European countries have been very low (between .5 and .8 percent) since the 1960's, the rate in the United States has been in excess of 2 percent. For the United States and United Kingdom, the excess of the growth rate of real *GNP* over the rate of labor force growth over this period has been modest (around 2 percent), while that for the other countries has been over 3 percent. This may imply a larger role for long-term aggregate demand shortages in the former pair of countries. In all of the countries, however, low demand growth and excess industrial capacity have characterized the 1970's, suggesting an important role for aggregate demand shortages in recent years.

Two other explanations of the recent unemployment remain. One structuralist conjecture concerns the role of youths (16–24 year olds) and females in the labor force. While the European countries, as a whole, have had zero absolute growth in the youth labor force from 1960–75, in the United States the youth labor force has grown by 82 percent over the period. Female labor force growth also varies substantially. For the Netherlands, the 1960–75 average annual growth rate of the female labor force was 1.9 percent; for Sweden, 2.3 percent; for the United Kingdom, 1.3 percent; and for the United States, 3.1 percent. For the United States, the female proportion of the labor force has grown from 33 to 40 percent since 1960. Thus, this structural case appears to be nonexistent for the European countries. And, for the United States, disaggregated data for the 1970's indicate that neither new workers nor reentrants have been major determinants of changes in measured unemployment (see Roger Brinner).

A second structural conjecture is that recent changes in the occupational and industrial composition of labor demand increments has been so skill specific as to reduce the possibility of a match with available labor supplies. To be sure, recent growth in labor demand has been of a rather unique industrial and occupational composition. For example, from 1965 to 1975, employment reductions in agriculture and manufacturing occurred in each of the European countries. The changes have not been nearly as severe in the United States. In the European countries, *all* employment growth has been accounted for by the tertiary sector. In that sector, occupational

demands are concentrated on service workers with high labor force elasticity and professional-technical workers with a low unemployment rate. While such a compositional shift might account for some of the apparent stickiness in the *response* of unemployment to demand growth, job specifications in the tertiary sector do not appear to be so narrow as to account for very much of the high *level* of measured unemployment. Again, the role of such structural imbalances would not appear to be substantial (see Ralph Turvey).

A final explanation for the current unemployment problem is that any unemployment rate (*U*) in the post-1970 period represents *less* excess labor supply than in the 1960's; that relative to earlier periods, current measured unemployment overstates the severity of both the recession to which it is often attributed and the social hardship which it implies. During the past decade, and especially in Western Europe, numerous legislative and bargained measures have altered the relationship between changes in the excess supply of labor (caused by shifts in either labor demand or supply) and *U*. Because the effect of these changes on the *U-ES* relationship may be either positive or negative, any net shift in the relationship depends on their relative strengths.

Consider first measures increasing the *U-ES* relationship. In Western Europe, one important change concerns policies relating to foreign workers. Since 1973, European Economic Community (*EEC*) countries have banned non-*EEC* labor recruitment. As a result, net emigration has fallen significantly, and foreign workers once in Western Europe have been reluctant to leave even though jobless. Unemployment would increase even if there were no increase in excess labor supply. There is some evidence that this had happened (see Joyanna Moy and Constance Sorrentino).

A second change inducing an increase in the *U-ES* relationship is the well-documented growth in generosity and coverage of work-tested income transfer programs. Since the net gain from employment is reduced by increased transfer generosity, covered workers have incentives to prolong job search (which is reflected in the measured duration of unemployment), to refuse to accept work except at higher offered wages, or to cease active labor market participation. Where the transfer benefit is accompanied by a work test, an increase in *U* is to be expected. It is estimated that from .3 to 1.2 points of the observed *U* in various countries is attributable to this factor (see Martin Feldstein, Dennis Maki and Zane Spindler, and Stephen Marston).

Another policy-induced increase in the *U-ES* relationship is the rapid increase in minimum wages in the Western European countries, particularly around 1969–70. For example, in the Netherlands, the minimum wage increased by about 50 percent, 1968–72, and more than doubled from 1966 to 1976, reaching about 70 percent of the median wage in 1977. This is likely to have seriously affected employment prospects for low productivity workers and new entrants in these countries. While the increase in minimum wages in the United States has been less than in Western Europe, the adverse affect of this legislation on observed youth and minority unemployment rates there is quite generally accepted (see Edward Gramlich; Terence Kelly).

Finally, while the changing composition of the labor force due to the rapid entrance of youths and women has been classified as a structural phenomenon, this could also be interpreted as an increase in the *U-ES* relationship. This compositional effect, associated with the higher turnover and shorter labor force duration of some classes of workers relative to others, has been well-documented (see George Perry, 1970).

Increased enterprise costs associated with hiring or releasing workers also alter the *U-ES* relationship. Recent legislation in the Netherlands, for example, requires firm wage liability for up to 2.5 years in case of employee lay-off or termination. In the

United Kingdom up to thirty weeks of severance pay can be required. As a result of these constraints, enterprises will reduce the variation over time in the size of the labor force, inducing reduced money wage elasticity—labor becomes a quasi-fixed factor of production (see Walter Oi). These programs will tend to reduce the *U-ES* relationship, although the resulting increase in labor costs also may generate effects similar to the minimum wage. The *U-ES* relationship is also reduced by recent substantial employment subsidies by Western European governments to enterprises designed to encourage labor retention (see Daniel Hamermesh).

Finally, there is the effect of the rapid increase in benefit levels and reduction in eligibility requirements of nonwork-tested income transfer programs. As with the work-tested programs, these changes increase the relative attractiveness of nonwork. However, unlike the work-tested programs, choice of participation in these programs implies a reduction in both employment and labor force participation, but not in measured unemployment. In effect, these programs have tended to become repositories for less productive (often older) workers who, in an environment with fewer constraints on enterprise hiring and firing decisions, would have been accounted as unemployed. Evidence of this change is the rapid growth in disability rolls in all of the countries studied during the last decade. For example, in the Netherlands the number of disability law recipients has grown from 215,000 in 1970 to about 530,000 in 1977. The large-scale training and education programs in Sweden, triggered by reductions in aggregate demand, have much the same effect (see Lennart Forseback). Given the constraints on enterprise hiring and firing decisions, both Western Europe employers and trade unions have supported the increased generosity of these programs. These legislated and bargained constraints on labor market performance probably have had a major impact on measured unemployment rates. However, because their impact on the *U-ES* relationship operates in both directions, it is difficult to determine their net contribution to the current unemployment problem.

III. Evidence on the Role of Structure and Measurement

Empirical estimates of Phillips curves and shifts in them are typically based on estimated money wage changes ($\dot{W}$) and U and involves some correction for price inflation (see Robert Flanagan; Perry, 1970). If U is a consistent indicator of excess labor supply (labor market tightness), observed shifts in Phillips curves are properly attributed to structural changes in labor markets, as described above. However, if U is *not* a consistent indicator of excess labor supply, observed Phillips curve shifts are attributable to *both* structural phenomena and changes in what the unemployment rate is measuring. The joint contribution of these two effects to observed shifts in Phillips curves is shown in Figure 1.

Assume P is the initial real relationship between wage rate increases ($\dot{W}$) and excess labor supply (ES). Given the initial *U-ES* relationship (M), a well-specified empirical estimate of the Phillips curve will yield $\hat{P}$ in the southwest quadrant. The effect of a structural change represented by a shift of P to P' on the estimated Phillips curve can now be observed—$\hat{P}$ shifts to $\hat{P}'$. This is the common understanding of one cause of Phillips curve shifts. The figure, however, also indicates that the estimated Phillips curve is also dependent on the *U-ES* relationship. Given P', then, a shift in this relationship from M to M' will also yield a shift in the estimated Phillips curve, from $\hat{P}'$ to $\hat{P}''$. Hence, estimated Phillips curves are dependent on both structural phenomena and the *U-ES* relationship.

To test the effect of both structural and measurement changes on estimated Phillips curves and implicitly on the unemployment problem, equations relating wage increases to unemployment and to consumer price

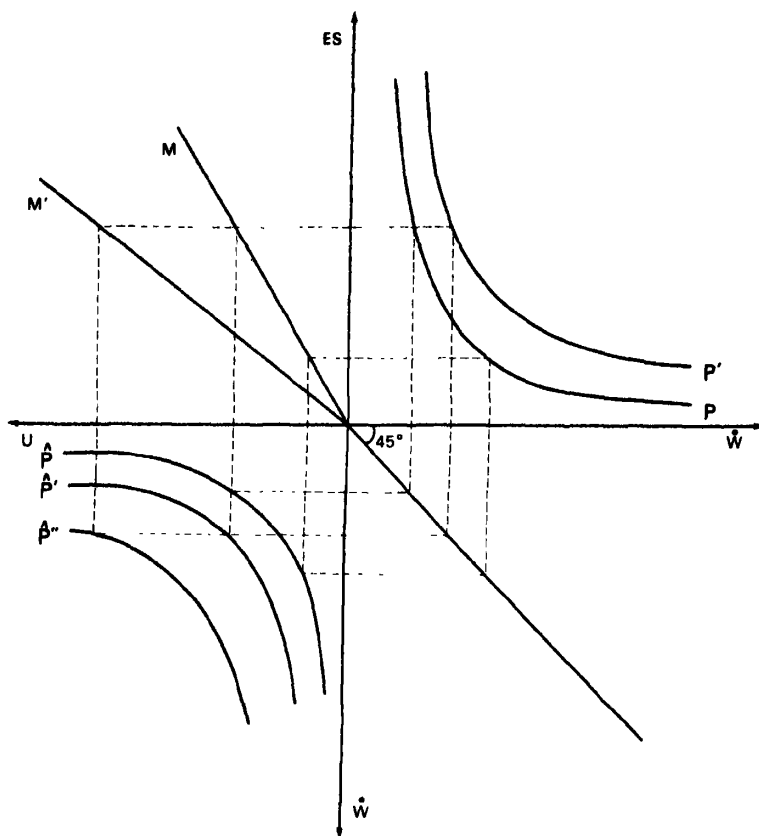



FIGURE 1

inflation were fit to pre- and post-1969 data for the four countries studied. Implicit in this procedure is the presumption that the post-1969 estimated Phillips curve is significantly different from the pre-1969 relationship because of the recent structural changes or changes in the *U-ES* relationship, or both. As shown in Table 2, in all cases a significant upward shift in the Phillips curve has occurred. Through the relevant unemployment ranges, the estimated $\dot{W}$ associated with any *U* has nearly tripled in the United Kingdom, nearly doubled in the Netherlands, and has increased by about 80 percent in Sweden and the United States.

The magnitude of these shifts is impressive, suggesting that some combination of structural changes or changes in the relationship of *U* and labor market tightness has occurred in all of the countries in recent years. And, if the limited support for the structural conjecture suggested above is accepted, the factors inducing an increase in the *U-ES* relationship—growth of work-tested transfers, altered policy on foreign workers, increases in minimum wages, and shifting labor force composition—must be viewed as exceeding those with the opposite effect and as playing a nontrivial role in explaining the current high level of unemployment in both Western Europe and the United States.

IV. On Unemployment Differences Among Countries

Overall, Sweden's performance in maintaining low measured unemployment

TABLE 2—WAGE INCREASE-UNEMPLOYMENT RELATIONSHIPS FOR THE NETHERLANDS, SWEDEN, UNITED KINGDOM, AND THE UNITED STATES, 1952–68 AND 1969–76

	Predicted Annual Rate of Wage Increase[a]							
Unemployment Rate (Percent)	Netherlands Pre-1969	Netherlands Post-1969	Sweden Pre-1969	Sweden Post-1969	United Kingdom Pre-1969	United Kingdom Post-1969[b]	United States Pre-1969	United States Post-1969[b]
1	9.2	20.6	9.2	21.8	6.7	17.8	15.1	1.5
2	7.6	14.8	6.2	12.1	5.3	15.9	8.0	5.0
3	7.3	13.7	5.2	8.8	4.8	15.3	5.7	6.1
4	7.2	13.3	4.7	7.2	4.6	15.0	4.5	6.7
5	7.1	13.1	4.4	6.2	4.5	14.8	3.8	7.0
6	7.1	13.0	4.2	5.6	4.4	14.7	3.4	7.3
7	7.1	13.0	4.1	5.1	4.3	14.6	3.0	7.4
8	-	-	-	-	-	-	-	7.4

[a]The brackets indicate the range of the unemployment rate during the period of estimation. Estimated from wage increase equations fit to quarterly time-series data on the unemployment rate in manufacturing, the consumer price index, and export and import prices. An appendix describing the estimated relationships is available from the author.

[b]The coefficient on the wage equations yielding this unemployment-wage increase relationship is not statistically significant.

in the 1970's is better than that of the other countries. The performance of the Netherlands and the United Kingdom lie at the other extreme. While several differences among the countries emerge from a perusal of labor market changes in each, such comparisons must be speculative at best. First, the key difference between Sweden and the rest of the countries appears to be the magnitude and effectiveness of the large-scale education and training programs for workers who would otherwise be unemployed. While these programs serve to maintain a very low U, a more comprehensive measure of unutilized or underutilized labor would undoubtedly diminish the strength of their performance.

Second, as between United States and the European countries, one important difference appears to be the relative growth in those factors causing an increase in the U-ES relationship. While minimum wages and the generosity of work-tested transfer programs have expanded in the United States, the magnitude of the change is much smaller than in Western Europe. Also, Western Europe has experienced the shift in policy on foreign workers. A second important difference concerns the magnitude of compositional changes in the labor force. The growth in the youth and female components of the labor force has been much more rapid in the United States than in any of the other countries studied. To the extent that structural factors are related to the high unemployment, they appear to be most severe in the United States.

Finally, even though substantial changes in the constraints imposed on workers, enterprises, and markets have occurred, and some structural imbalances have developed, it is difficult to conclude—especially in the case of the United Kingdom and the United States—that a lack of aggregate demand is a trivial cause of the current unemployment problem. Such a demand shortage, moreover, may well have its source in policy-based incentives for capital-labor substitution or for enterprise flight in small economies, related in part to rapid increases in taxes on the wage base. In some cases (for example, Belgium, Italy, and the Netherlands), the combined employer-employee tax rates are now in excess of 50 percent (see Willem Driehuis).

REFERENCES

R. Brinner, "The Death of the Phillips Curve

Reconsidered," *Quart. J. Econ.*, Aug. 1977, *91*, 389-418.

W. Driehuis, "Capital-Labor Substitution, Technology, and Employment," mimeo., OECD, Paris 1977.

M. Feldstein, "Lowering the Permanent Rate of Unemployment," Joint Economic Committee, U.S. Congress, Washington 1973.

R. Flanagan, "The U.S. Phillips Curve and International Unemployment Rate Differentials," *Amer. Econ. Rev.*, Mar. 1973, *63*, 114-31.

Lennart Forseback, *Industrial Relations and Employment in Sweden*, Stockholm 1976.

E. M. Gramlich, "Impact of Minimum Wages On Other Wages, Employment and Family Incomes," *Brookings Papers*, Washington 1976, *2*, 409-61.

D. Hamermesh, "Indirect Job Creation in the Private Sector: Problems and Prospects," in John Palmer, ed., *Direct Job Creation*, Washington forthcoming.

T. F. Kelly, "Two Policy Questions Concerning the Minimum Wage," work. paper no. 3608-4, Urban Inst., Washington 1975.

D. Maki and Z. Spindler, "The Effect of Unemployment Compensation on the Rate of Unemployment in Great Britain," *Oxford Econ. Pap.*, Nov. 1975, *27*, 440-54.

S. Marston, "The Impact of Unemployment Insurance on Job Search," *Brookings Papers*, Washington 1975, *1*, 13-48.

J. Moy and C. Sorrentino, "An Analysis of Unemployment in Nine Industrial Countries," *Mon. Labor Rev.*, Apr. 1977, *100*, 12-24.

W. Oi, "Labor as a Quasi-Fixed Factor," *J. Polit. Econ.*, Dec. 1962, *70*, 538-55.

G. Perry, "Changing Labor Markets and Inflation," *Brookings Papers*, Washington 1970, *3*, 411-41.

———, "Determinants of Wage Inflation Around the World," *Brookings Papers*, Washington 1975, *2*, 403-35.

R. Turvey, "Structural Change and Structural Unemployment," *Int. Labor Rev.*, Sept./Oct. 1977, *115*, 209-15.

Unemployment Problems and Policies in Less Developed Countries

By HENRY J. BRUTON*

The conventional classification of an entire population as employed, unemployed, or outside the labor force raises many questions in any society. In many of the less developed countries of the world the issues raised by such a classification are even more numerous and are of greater relevance for policy than is the case in Western Europe and the United States. I shall discuss the nature and content of unemployment in the less developed countries and suggest some of the implications for employment and development policy of those observations.

In seeking to define and measure employment and unemployment in any society, we are forced to make a large number of rather arbitrary assumptions. Most employment/unemployment data are collected by surveys, and in forcing everyone into one of three categories, it is almost inevitable that we lose a lot of information, resulting in an unrevealing and even misleading picture. Generally a person is listed as employed if that person had worked for pay or profit to himself or his family on at least one day (sometimes even less, sometimes more) during the reference period, usually the week immediately preceding the survey. A lower age limit is often applied so that persons under 14 or 15 years of age are automatically defined as outside the labor force. Some countries have an upper age limit as well. Allowance is usually, but not always, made for people on sick leave or vacation. Similarly, a person is classified as unemployed if that person had not worked at all during the reference period and was either actively seeking work or was available for work at the going wage.

Presumably, the reference period is meant to be average or typical, and the figures obtained would then tell us how many people were employed and unemployed in this typical period, according to the definitions just stated. An annual or semiannual survey taken over an extended period would then reveal what was happening to the two variables over time.

Rates of unemployment in less developed countries based on this approach and these definitions often do not appear excessively high, especially for the economy as a whole. Rates are almost always higher in urban centers than for the entire country. Figures in the 3–8 percent range are most frequent, although some countries and areas, Sri Lanka and Java for example, much exceed this range.

There are, however, a number of reasons why such results are less useable in developing countries than they are in the more developed world. They certainly do not identify and measure what most of us think of as the major problem of these countries. Consider a few of the difficulties.

The most obvious difficulty arises from the fact that own-account workers constitute, in almost all developing countries, a much larger percentage of the labor force than in the more developed countries. In Tanzania less than 40 percent of the labor force are classified as "employees," in India less than 20 percent, Indonesia about 32 percent, Korea 37 percent, and so on. In the United States and the United Kingdom over 90 percent of the labor force are classified as salaried employees or wage earners. Own-account workers fit the conventional classifications much less well than do wage earners. A shop keeper who has only a couple of customers per day, a rickshaw taxi that carries only two or three passengers a day, a farmer who has only enough land to keep him fully busy now and then are all employed by the usual definitions, but that fact is not very comforting to us. In some instances the own-account

*Williams College.

worker earns an average livelihood. In other instances, of course, such workers are terribly poor, and have made themselves a job simply because they cannot afford the luxury of unemployment.

A similar point concerns the frequency with which one encounters people with more than one job. If an individual has three jobs, and loses one of them, should the person be classified as unemployed?

The labor force notion is more ambiguous for developing than for developed countries. An extreme example is found in a 1964 Socio-Economic Survey of Indonesia. In that survey about 2 million women whose "main occupation" was housekeeping were all classified as "outside the labor force" even though 1.5 million of them worked twenty or more hours per week in jobs that ordinarily are counted as employment. On the other hand 1 million women who worked in activities that were counted as employment for less than fifteen hours were classified as "within the labor force" because they did no housekeeping. In the Sudan, women often do most of the farming of family lands as men migrate to find work and are frequently away from home for a year or more at a stretch. Yet most data would show that the men did the work, and the women were "housewives." Another example has to do with education. In West Germany less than 10 percent of the 18–21 age group are classified as full-time students. In India over one-quarter of that age group are full-time students. The explanation surely is in terms of employment opportunities.

Also, of course, many people, women especially, would accept a job if it were near at hand or allowed time off to do household chores or whatever, but no such jobs are available so no search is made. To classify these people as outside the labor force is not very revealing. The International Labour Organization's employment report on Kenya classifies the unemployed as those "with zero income who are seeking work." This is misleading on both counts. People with zero income do not survive, and those who want work in a real sense may not search. Thus the distinction of being in or out of the labor force is an unhelpful categorization in many countries. It is especially unhelpful in rural areas where most of the population of developing countries live.

Government policy often hides unemployment from the survey enumerator. In Egypt, for example, the government guarantees every university graduate a job. Since university education is virtually free to the individual, and since a job is promised, university enrollments are enormous. The quality of education is low and probably has been falling in recent years. The result is a large and increasing number of government employees who, for all practical purposes, do nothing in the way of productive work. Yet unemployment figures for Egypt in no sense capture that fact.

In other countries government policy has in the past made unemployment possible. In Sri Lanka family allowances from the government were such that young people just entering the labor force could be very particular about the job they would take, and hence were classified as unemployed. At the same time Indians were imported to perform a number of jobs in agriculture. Many other examples are available.

These examples illustrate the importance of the means of support available to the unemployed. Unemployment, as conventionally defined, requires that some means of support be available to the individual, and dependents. That less developed countries rarely have well-organized formal means of providing such support often means that the very poorest people are the working poor rather than the unemployed. It also means that public "employment" is often used as a means of providing "unemployment" insurance over a more or less indefinite time period.

One more specific issue emerging from the characteristics of many less developed countries is worth a brief comment. A sector exists which pays wages very much higher than the average in the economy as a whole. This high wage, not surprisingly, attracts applicants. In some instances the job seekers come from other areas of the

country, in other instances the applicants are nearby residents. In both instances they in effect spend their time standing in line at the gate of the high wage paying activity in order to maximize the probability of being employed in this particular activity. How long they can stand in line depends on their means of support. Several studies are available to show that an individual member of a family is often supported by other family members in order that he or she may spend the time seeking the high wage jobs. In many instances household income must increase in order to provide the means of support for a family member to become unemployed. These people then are unemployed by the usual definitions, but such unemployment has different implications—social and economic—from that which usually concerns us. It is due to the fact that the characteristics of the developing countries produce a sector or activity in which wage rates are very much higher than the economy-wide average. More general perhaps is the existence of a large wage differential between rural and urban areas. This of course also pulls people into urban areas to look for jobs, even where lower paying employment opportunities do exist in the rural areas. This is a kind of distortion that is rarely observed in the more highly developed country.

The implications of this situation vary. A substantial part of urban wage earnings are remitted to rural areas. The percentage remitted seems to be higher, the lower are wage earnings. In this event the sharing arrangements alleviate some of the inequities and other problems created by high urban wages. However, as ties with the rural areas break down and as modernization proceeds, this sharing arrangement begins to fail and the situation becomes more difficult.

These examples suggest some generalizations about the sources of the difficulties of using conventional definitions and classifications to illuminate the unemployment problem in developing countries. In the first place, the large proportion of own-account workers can mean many things. It does mean, in varying and unidentifiable degrees, that people are trying to eke out an existence simply because they have no alternative means of work or support, but it can also mean that many people have found a productive and remunerative role to play in the society as it exists. Second, social and economic organization in many developing countries results in the line between employment and nonemployment, and participation and nonparticipation in the labor force being not merely blurred, but in a very real sense nonexistent for many people. This is especially evident with respect to women in rural areas, but is not limited to this group. Third, the general inability of most developing countries to maintain any sort of formal unemployment or welfare payment arrangement often means that employment is a means by which the unemployed are sustained. Government employment and personal servants are perhaps the most common, but not the only, examples. Finally, a particular set of market distortions has created an inducement to become unemployed in order to exploit or seek to exploit that market distortion. Of greatest relevance in this respect is the wage differential between the rural and urban areas.

The results of employment surveys of the kind referred to above do not, therefore, tell us a great deal about the two things of greatest concern to the policymaker: the availability of labor for employment in various activities and the welfare effects of unemployment. It is worth repeating that many of the unemployed (in the survey sense) are living better than those classified as employed. This can mean that the unemployed with means of support are often less available for new activities than are the very poor who must have some kind of a job.

Consider now an argument or two related to policy matters. The general conclusion that labor in the less developed world generally has very low (possibly zero) marginal productivity led to the argument that the country should seek a very high rate of capital accumulation to supply inputs complementary to the labor and

hence to raise the productivity of the latter. This kind of notion directed the policymaker's attention toward factors affecting the rate of capital accumulation in the modern sectors. It also tended to encourage the policymaker to think of the traditional sector as some sort of bottomless reservoir in which those who could not be absorbed into the modern sector would rest undisturbed until the capital accumulation had proceeded long enough to pull everyone into the modern sector. Although rates of capital accumulation in the 1950's and 1960's were often respectable, results in terms of poverty alleviation and modern sector employment have almost always been disappointing.

A great deal of documentation is now available to show that many countries so distorted incentives and signals in favor of capital that employment was heavily penalized. It is, I think, safe to say that alternative policies were available that would have resulted in a significantly higher rate of growth of employment and more success in the alleviation of the most severe forms of poverty than has been achieved. At the same time, it also seems consistent with the general picture now emerging that to limit the attack on our objectives simply to seek higher and higher rates of capital accumulation does not carry much promise of full success, even with improved policies. There are country experiences that dispute this, for example, Taiwan, Korea maybe, but it seems true in general.

The other source of the failure to which I wish to devote more attention has to do with the notions of employment that I have been discussing. President Nyerere of Tanzania is reported to have said that, "In the old Africa everybody worked." He seems to have meant that everyone performed a role or assignment that justified a claim on the society's available output. In this environment there was little inequality and little unemployment in the eyes of the community. A.K. Sen notes that many urban workers return to the farm during the busy season not so much because more labor as such is needed, but in order to establish a claim on a share of output. We are observing a breaking down of the relationships that prevailed in the traditional societies at the same time that we have failed to establish a generally effective alternative to the traditional arrangements. A study of the data produced by employment surveys placed alongside a more complete description of how individuals spend their time and how output is in fact distributed illuminates this failure.

A couple of examples may be useful. A land reform program that breaks up large estates may result in a reduction in measured employment, while household standards of living improve as all family members contribute in one way or another. This seems to have been the case as European owned and operated farms in East Africa (and elsewhere) were broken up into small holdings. In several countries (for example, Tanzania and Indonesia) concentrated surveys in limited areas almost invariably show larger, sometimes much larger, numbers occupying remunerative roles than do routine employment surveys. Such roles often demonstrate a specific degree of entrepreneurship and ingenuity in adapting to a social and economic environment. These activities invariably fit the environment more suitably than do those that are actively encouraged by the government and many outside agencies.

I might say something rather extreme just to help make the point: the employment/unemployment/outside the labor force classification encouraged the view that we could, by capital formation in the modern sector, create enough new productive jobs to relieve poverty and to create a role, a place for everyone in modern sector activities. This simply has not worked. As we look more closely at the activity of labor in the developing countries, additional possibilities can be seen. In general we may understand how to improve upon, how to build upon those traditional arrangements with respect to the use of labor and the distribution of output that seem to work well, rather than to ignore or penalize them to try

and build something quite alien to the society that is intended to absorb in some way or other all members of the labor force.

Revealing ideas along this line are found in some of the International Labour Organization's studies on employment, especially those on Kenya and the Philippines. The examination of what the Kenya report calls "the informal" sector shows that activities in this sector—often penalized and rarely helped by official policy—provide a satisfactory source of livelihood for many people at very low rates of investment. The role that rural, nonagricultural activities can play (and have played here and there) in providing work activities of a rewarding sort is becoming increasingly clear. The study of sharing arrangements is equally important, if for no other reason than to avoid breaking down an existing sharing arrangement before employment or other, new means of support are available.

These observations suggest that we avoid relying on the three-way classification mentioned earlier. In its place attention would be focussed on a variety of questions intended to illustrate how the adult members of the household spend their time, how and to what extent the activities performed by household members are rewarded, and the availability from government and nongovernment sources of money and income in kind not in payment for services rendered. Especially important are data indicating how adults spend their time and why they spend it in the way they do. Such data would help us understand how worker activity "fits in," how it contributes to the way the system operates, and how this contribution could be enhanced within a given environment. Such understanding would also help us to identify pressure points or change inducing points that could provide places to modify the arrangements.

Data indicating sources of income—those based on work, those based on recognizable personal claims, those based on social claims, etc.—will obviously help us to understand more clearly welfare issues. Also they would indicate something about labor mobility, about incentives and response to incentives, about sharing arrangements, about social relationships that affect economic behavior, about the legitimacy of certain practices—often deplored by outsiders—in a given context.

One-dimensional series such as employment or unemployment are extremely convenient, and I do not really think that they should be abandoned completely. I do think however that trying to put the kind of information just referred to into a framework, and trying to find the rationale underlying its existence, would increase our understanding of the societies in the less developed world. It might also facilitate the design of policies that would build upon existing institutions and arrangements. It may also help to lead us away from policies that destroy these arrangements without really offering anything to replace them.

DISCUSSION

NANCY SMITH BARRETT, American University and Urban Institute: Since 1971 inflation and unemployment have exceeded postwar averages in Europe and the United States. Robert Haveman attributes this phenomenon to a change in the relationship between measured unemployment and excess labor supply. While this factor may explain some of the increase, surely a more important explanation is the simultaneous occurrence of strong inflationary shocks from higher food and fuel prices and attempts by workers to protect their real wages by seeking cost-of-living adjustments, combined with restrictive macro policies to ward off balance-of-payments crises and to defuse political opposition to inflation. It is true that higher unemployment has coincided with higher inflation. However, this says nothing about the tradeoff, that is, what would happen to the inflation rate if aggregate demand policies pushed unemployment lower; nor does it mean that the labor market has become tighter for a given unemployment rate.

Before 1971, unemployment rates in Europe were typically in the 1 to 2 percent range, while in the United States they ranged from 4 to 6 percent. Although some of this difference resulted from a discrepancy between survey methods and definitions, most of it was related to other factors. For instance, there is less labor mobility in Europe than in the United States. This means less frequent job change and also a lower incidence of unemployment. Also, in the pre-1971 period, the Europeans were operating closer to potential than was the United States.

After 1971, unemployment rates rose markedly in the United States and Europe. But two distinct phases must be identified. The first is 1971–73 when European rates rose to around 2.5 percent and the *U.S.* rate to around 5.5. This unemployment resulted from a sharp drop in aggregate demand. In the United States the recession was due to restrictive macro policies designed to shake out the inflationary legacy of Vietnam, and the downturn in the United States was clearly the principal impetus for the recession in Europe. Wage inflation continued both here and abroad due to inflationary expectations. However there surely cannot be any question that the higher unemployment in those years was due to a lack of effective demand—not to structural factors.

The second phase began with the food and fuel price explosions of 1973–74. Despite attempts to apply incomes policies in various countries, it was impossible to hold down wage increases in the face of a 400 percent increase in the price of imported oil and a 25 percent jump in food prices. All of the countries except for Sweden quickly enacted restrictive macro policies both to hold down inflation and to stave off balance-of-payments crises. The crucial role of the restrictive policies followed by West Germany surely should also be mentioned. In this period, labor interests fought erosion of real wages by seeking catch ups to the spiralling food and fuel prices. Simultaneously, unemployment rose as a result of restrictive policies. While this shows up statistically as a worsening of the Phillips relation, it was due to the contemporaneous effects of separate influences on wages and unemployment rather than an indication that the labor market had become tighter for some given unemployment rate.

Having objected to Haveman's conceptualization of the problem, I am afraid I shall also have to take issue with his explanation. He rejects the inadequate demand explanation on the basis of a rough calculation of labor demand and supply growth for the period since 1960. Since it is an annual average calculation, the experience of the most recent years is swamped by the buoyant performance of the 1960's, when labor was very scarce in Europe and when the United States was involved in a major Asian military operation. Extrapolating from a trend dominated by the 1960's is hardly a way to characterize labor market demand conditions in the period since 1971, nor an appropriate framework for analyzing the recent rise in unemployment.

Haveman then moves to structural expla-

nations of unemployment. He rejects demographic explanations for the European countries, as well as the notion that there is a mismatch between the skill requirements of various jobs and those of the labor force. By process of elimination, he reaches the conclusion that any recorded unemployment rate in the post-1970 period represents less excess labor supply than in the 1960's. Nowhere does he test this hypothesis against the others, so at best, it reflects his considered judgment. He cites three major factors contributing to this change: policies relating to foreign workers; an increase in work-tested transfers like unemployment insurance; and higher minimum wages.

I cannot dispute the contention that these policies have added somewhat to measured unemployment, but they seem negligible compared to the effects of restrictive demand policies. The fact that Sweden has kept unemployment so low by maintaining (until recently) fairly high levels of aggregate demand combined with severe restrictions against private-sector layoffs and large scale public employment and training programs suggests that foreign workers, unemployment insurance, and rising minimum wages only cause higher than normal unemployment when people cannot get jobs.

If a government in a welfare-capitalist state wants to pursue restrictive demand policies to avert potential balance-of-payments crises (as in Europe) or to defuse political opposition to inflation (as in the United States), some sort of income- or work-conditioned transfers must be made available for those out of work. To the extent that the side effect of such programs is some small rise in measured unemployment over what it would have been in the absence of such programs, one can conveniently blame them for all or most of the unemployment. However it is rather like blaming the cow for leaving once the barn door has been left open.

Many changes are occurring in the labor markets of industrial countries, and some of these are rooted in common factors that make international comparisons interesting and useful. Some of the influences described in Haveman's paper are part of the process of change. But it is not legitimate to single out any one of them for special emphasis without the weight of empirical evidence.

My own view is that unemployment in the United States and most of Western Europe is the result of rather unimaginative responses to oil deficits and inflationary fears rather than major structural changes in labor markets.

Multiple Motives, Group Decisions, Uncertainty, Ignorance, and Confusion: A Realistic Economics of the Consumer Requires Some Psychology

By JAMES N. MORGAN*

How can we have more realistic theory and research about economic behavior without an explosion of variables or a retreat into crass empiricism? One method is to start with the main and most obvious ways in which a simple one-dimensional utility-maximization theory fails to reflect reality. I shall discuss five that I think are most crucial:

1. Multiple motives—A person with a set of needs or desires faces a variety of products or services, each with multiple attributes. The utility of any one product or service is actually a complex function of its attributes.
2. Group decisions—In the family and probably even in the firm, there are several individuals with differing patterns of needs, trying to achieve some consensus about what to do. Elements of a power struggle mix with altruism. Communication may fail.
3. Uncertainty—Most decisions involve commitments about an uncertain future. The length and size of such commitments can be expected to vary with the degree of uncertainty about the future.
4. Ignorance of facts—Knowledge about the range of alternatives and their actual prices and qualities is not readily available to most of us without work and expense. Decisions must be made about investment in information, and there are problems of decision making when the facts are *not* all known.
5. Confusion—Economic decisions require an understanding of economic principles to know which facts are needed, and what to do with them. The great policy issue of whether consumers need more facts or more protection hinges on whether they would know what to do with the facts if they had them; whether consumers are not just informable, but educable!

*Research scientist, Institute for Social Research, and professor of economics, University of Michigan.

I. Multiple Motives

Theoretical writings by Kelvin Lancaster (1968, 1971) have discussed the implications of the fact that any commodity has a number of characteristics which must be merged with a number of needs of the individual to produce a desirability (utility?) indicator. There have been some statistical attempts to tease out shadow prices of attributes of cars or houses or jobs from the market prices of those with varying combinations of attributes. This puts a heavy burden on the data in view of various supply conditions and the need for assuming all or large blocks of consumers have the same tastes.

What seems to be missing in all this is sufficient attention to the differing motives or needs of the consumers. Tibor Scitovsky's book, *The Joyless Economy*, focuses on the need for novelty and stimulation as well as comfort and satiety, but surely we would want to add desire for power over others, affiliation with others, and achievement against obstacles as also motivating much of our behavior.

Perhaps the most useful paradigm for making use of motives is that of John Atkinson who argues that there is an incentive level of each need, depending on how much it has already been satisfied (declining marginal utility), a basic level of importance of that motive, and a subjective

probability that some course of action (acquiring a product?) will indeed produce some of that sort of satisfaction.[1]

If we are primarily interested in *change* in behavior, then we may want to focus on two things: the way experience changes people's notions about the probability that some particular product, activity, or choice will produce satisfaction; and the extent to which aspiration levels tend to be satiated on the one hand, or to rise with each achievement on the other.

For example the life insurance provisions of the Social Security System did not eliminate private life insurance, and some argue that it made adequate provision possible and actually stimulated additional private insurance to make the coverage really sufficient. A study of the impact of private pensions when they were initiated, largely involuntarily, indicated that if anything they raised aspirations for a truly adequate retirement income and increased private saving. (See George Katona, 1965.)

Focusing on personal motives and perceived paths for satisfying them may well be a more fruitful approach to consumer behavior than focusing on attributes of products. People's perceptions and subjective probabilities surely change more rapidly than product characteristics, and people's aspiration levels may also change under the impact of experience, success, failure, or even persuasion.

II. Group Decision Making

If we have done little on how multiple motives combine with multiple product attributes and with some probabilities to generate expected utilities, our ability to deal with group (family) decision making is even more deficient. Kenneth Arrow proved some time ago the impossibility of any simple social welfare function, and the same problems arise with a family attempting to develop a family set of preference orderings.

Since families do make decisions, a majority of them sufficiently satisfactory so that the family stays together, we might ask whether we can develop some notions about their methods. There are possible contributions from psychology on power, conformity and altruism, from sociologists on roles and role expectations, and from economists on multilateral trade and bargaining. Market researchers have asked directly who made the decision, and have found the usual ego bias in the answers, plus a tendency to conform to social roles as to who is expert in what—women on style, color, and the household, men on mechanics and the car. A study by Elizabeth Wolgast that looked at who could predict what the family would do found the women predicted better, even in the male areas.

We could surely use more studies of communication and consensus in the family—who knows what others want, does actual agreement go along with perceived agreement? And where there are differences, or disagreements, we could perhaps be studying the relative role of power and affiliation in reaching acceptable compromises. We do not, of course, have a good theory to handle decision making when affiliation motives dominate. Indeed, one of C. S. Lewis' funniest episodes in the *Screwtape Letters* is one where everyone tries to do what the others want, and they all end up angry. Since the family is the main source of support for otherwise dependent members of society, overwhelmingly more important than social security and welfare, its stability is an important public policy issue.[2]

III. Uncertainty

For more than a quarter of a century now, the main work on the actual effects of

[1]See Atkinson, Atkinson and Norman Feather, and Atkinson and Joel Raynor.

[2]For an assessment of the quantitative importance of change in family composition in accounting for the variance in change of status, see Greg Duncan and the author (1977). And for a study of factors affecting changes in family composition, see Duncan and the author (1976). For estimates of the value of intrafamily transfers, see Nancy Baerwaldt and the author (1978). For attempts to clear up the conceptual confusion about social security, see the author (1976, 1977).

uncertainty on the consumer has been that of Katona. Coming to this problem with a background in psychology *and* economics, Katona decided that since the crucial economic behavior for economists was spending and committing funds, and since the major dynamic explanatory variable was probably uncertainty about one's personal financial future and about the country's economy, one should start investigating the link between events, mass attitudes, and mass consumption expenditures.

It was clear years ago that changes in consumer investment expenditures were large, dramatic, and more difficult to predict than changes in business investment expenditures. It tells us something of the difficulties of using psychology in economics in that it has taken Katona a quarter of a century to convince economists that there can be mass changes in willingness by consumers to spend and make commitments, as events and people's perceptions of them alter their confidence in their own and their country's economic future. It seems such a straightforward proposition, such a relatively simple theory, and so crucial to dynamic economic analysis and forecasting. And it did not multiply the explanatory variables inordinately. A substantial accumulation of empirical evidence over many years shows that occasionally at crucial times there are mass changes in attitudes followed by changes in levels of consumption expenditure. The accumulated information on the national and personal events that preceded and accompanied those changes in attitudes allows the development of theory about how people see and interpret their environment, and how they learn from history as each new event impinges on a revised set of information and beliefs.

Since aggregate (average) levels of consumer optimism predict individual spending behavior better than the individual's own optimism, we clearly need a theory about the way perceptions of other people's attitudes affect our own behavior.

IV. Ignorance

It is not merely uncertainty about the future that inhibits spending and affects its content. There is ignorance (or uncertainty about the facts), and also confusion about how one uses facts to make decisions on economic choices. The field of consumer economics is full of admonitions to consumers to stop doing irrational things, meaning failing to get facts or to use them properly.

Such behavior may not be irrational, however, if information is expensive to secure, relative to the possible gains. There is a growing theoretical literature in economics about optimal information-search strategy in both purchasing and job searching, though few empirical studies of actual behavior.[3] Of course, the economic theory of a market economy shows how it can approach optimality under a set of assumptions, a crucial one being that individual actors are informed.

We have proposed a continuous telephone-interviewing and feedback system for a local community that would test the economic feasibility of such a local information service, and at the same time assess its effects on market functioning.[4] Probability samples, various adjustments for biases, and the law of large numbers should allow us to bypass any need for precise measurement of quality of service. We argue that quality is whatever leads to satisfaction, for the most part, and that knowing the fraction of the customers of each competitor who are satisfied with the quality/price is the kind of information most of us have been using anyway.

Given what I have said about the inhibiting effects of ignorance of market facts, my guess would be that the main effect of such a system would be to expand the total demand for services toward the best vendors, while encouraging the others to improve. (Many of the deficiencies may well be unknown to the management.) Better information about the quality/prices of local repair services may encourage repairs and

[3]For studies of deliberation in purchasing, see Eva Mueller and Katona and Joseph Newman and Richard Staehlin. For studies of job search, see Harold Sheppard and A. Harvey Belitsky, and Mark Granovetter.

[4]So far it is only an unfunded proposal and some pilot work. See E. Scott Maynes et al.

conserve resources by postponing replacement.

Perhaps the most exciting thing about such an experiment is that it can have two alternative possible outcomes. First, the correlation between price and quality can improve so rapidly that no one would be willing to pay for the information about the remaining differences. In this case, the social benefits of information are obviously large, widely spread, and justify subsidy out of public funds. Second, the improvement may be slow, erratic, and subject to changing conditions among the vendors, so that it remains valuable to users to have the information. In that case, a local consumer information service is an economically viable system, and the design is exportable to other communities.

V. Confusion

I come now to the final area where economic theory needs to be supplemented both by psychological theory and by empirical research if it is to be useful; namely, how people apply the insights of economic analysis to ordinary choices, or fail to do so. There is abundant evidence that consumers lack the economic understanding and problem solving abilities to act in a way which economic theory says is optimal, given their goals. I have labeled this situation "confusion," because people are confused as to which facts they need and what to do with the facts in making decisions. The great policy debate as to the relative role of consumer information versus more consumer protection by the government hinges on whether the consumers are not just informable but also educable as to what to do with information.

For example, the payoff to finishing high school, it turns out, is the sum of a set of expected values, each the product of the payoff to an alternative future opened up by the high school diploma, times the probability that that will be the future path. Many of us don't really believe in probabilities, much less in *adding* the expected values of alternatives. And the payoff to shopping one more store turns out to depend on an estimate of price/quality variability that must be continuously adjusted for each new bit of information. Do we expect consumers to be Bayesians as well as probabilists?

There is some interesting research going on in information processing, but the most important problems of the consumer are not information overload, but lack of any cognitive structure.[5] We need empirical research here. It may be that consumers have other motives or constraints. James Duesenberry's famous statement reminds us: Economics is all about how people make decisions, Sociology is all about why they don't have any decisions to make. Indeed, social pressures, reference groups, acceptable roles, and the inevitable course of the family life cycle may dominate many decisions.

Katona (1975) has summarized much of what we know about the psychology of economic behavior. People are ignorant, but not dumb; they lack theory and insight, but have rules of thumb that often save them; they can be led by the media and other devices, but not very far away from what satisfies and pleases. Their behavior is for the most part meaningful to them, even though it may seem irrational by normative standards of deductive economic theory.

This may explain why studies of shopping and information-getting make consumers seem so far from optimum behavior as the economist defines it. They may be relying on information already half processed, less specific but more useful and cheaper to get. The readers of *Consumer Reports* (*CU*) over the years repeatedly insisted on rankings, and "best buy" indications, inferring that they are willing to trust *CU*'s criteria weights and save time and energy. An early study by Mueller and Katona indicated circumspect and deliberate shopping for sports shirts, and less shopping than expected for major appliances, presumably because there was more variety of style and more chance for regret with the former.

[5]Most of this work is coming out of Purdue. See Jacob Jacoby, Robert Chestnut, and William Silberman; Jacoby, Jerry Olson, and Rafael Haddock; and Jacoby, George Szybillo, and Jacqueline Busato-Schach.

It is entirely plausible that one result of the combination of ignorance, uncertainty, and confusion is inhibition of action entirely. Some years ago a research project provided a sample of consumers with market and quality information, and another sample with exhortations to buy wisely, and then went back to see which group seemed to be buying most wisely in terms of the price/quality information made available to one group. The results were clear—the group with detailed information *did* not buy noticeably "better," they simply bought *more*.

It is change in behavior that matters; hence the relevant psychological theory is that of perception and/or learning, combined with a theory of motivation. How do people perceive changes in their environment; what cognitive insights lead them to infer what they should do about it; and what do they learn from the results of their past behavior? This kind of dynamic question is hard to investigate. There is some panel study analysis (see Greg Duncan and Daniel Hill), and I have proposed a study that will ask people how their choices affected one another as each person settled into a job, a marriage, a spouse's job, a decision about children, and a location.

I also once proposed some research on why people do not use the logical processes the economists say they should. It was suggested that in a sequence of questions we could lead a respondent through the "proper" considerations and inferences, and then see whether he came to the expected conclusions, rejected them because of other considerations, totally refused to accept the logic, or understood everything but admitted making actual decisions much more casually.

VI. Summary and Conclusions

When there are many complicated problems, we need a research strategy. Such a strategy combines guesses both about potential results (productivity of the research) and about the value of the conclusions. In economic behavior, it is mass dynamics that matters most to the nation, not interpersonal differences that cancel out. We also want results that bear on policy issues, and may affect major government programs or policies. My own judgment is that pushing faster on the study of the effects of uncertainty and of confidence and optimism on willingness to spend and make commitments by consumers ranks first. Then would come research on the effect of improved simple information on consumer behavior and market functioning. Third, we need to find out how to educate people about economics so they can solve their own economic choice problems effectively. Fourth, we may want to know how families solve their joint decision problems, and lastly, how complex sets of wants combine with attributes of products in an uncertain world to affect people's desires.

REFERENCES

John Atkinson, *An Introduction to Motivation*, Princeton 1964.

——— **and Norman Feather,** *A Theory of Achievement Motivation*, New York 1966.

——— **and Joel Raynor,** *Motivation and Achievement*, New York 1974.

N. Baerwaldt and J. Morgan, "Trends in Intra-family Transfers," in Lew Mandell et al., eds., *Surveys of Consumers 1971–72*, Ann Arbor 1973.

J. Duesenberry, "Comment," in *Demographic and Economic Change in Developed Countries*, Universities-Nat. Bur. Econ. Res. conference series, Princeton 1960.

G. Duncan and D. Hill, "Attitudes, Behavior, and Economic Outcomes: A Structural Equations Approach," in Greg Duncan and James Morgan, eds., *Five Thousand American Familes*, Ann Arbor 1975.

——— **and James Morgan,** *Five Thousand American Families:Patterns of Economic Progress*, Vol. 4, 1976; Vol. 5, 1977; Vol. 6, 1978, Ann Arbor.

Mark Granovetter, *Getting a Job*, Cambridge 1974.

J. Jacoby, R. Chestnut, and W. Silverman, "Consumer Use and Comprehension of Nutrition Information," *J. Consumer Res.* Sept. 1977, *4*, 119–28.

———, **J. Olson, and F. Haddock,** "Price,

Brand Name, and Product Composition Characteristics as Determinants of Perceived Quality," *J. Appl. Psychol.*, Dec. 1971, *55*, 570–79.

———, **G. Szybillo, and J. Busato-Schach,** "Information Acquisition Behavior in Brand Choice Situations," *J. Consumer Res.*, Mar. 1977, *31*, 209–16.

George Katona, *Private Pensions and Individual Saving*, Ann Arbor 1965.

———, *Psychological Economics*, Amsterdam 1975.

Kevin Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Apr. 1966, *74*, 132–57.

———, *Consumer Demand: A New Approach*, New York 1971.

C. S. Lewis, *Screwtape Letters*, New York 1944.

E. S. Maynes et al., "The Local Consumer Information System: An Institution-to-Be," *J. Consumer Aff.*, Summer 1977, *11*, 17–33.

J. Morgan, "An Economic Theory of the Social Security System and Its Relation to Fiscal Policy," in George Tolley and Richard Burkhauser, eds., *Income Support Policies for the Aged*, Cambridge, Mass. 1976.

———, "Myth, Reality, Equity and the Social Security System," *Econ. Outlook U.S.A.*, Autumn 1977, *4*, 58–60.

———, "Intra-Family Transfers Revisited: The Support of Dependents Inside the Family," in Greg Duncan and James Morgan eds., *Five Thousand American Families*, Vol. 6, Ann Arbor 1978.

E. Mueller and G. Katona, "A Study of Purchase Decisions," in Lincoln Clark, ed., *Consumer Behavior*, Vol. 1, New York 1954.

J. Newman and R. Staehlin, "Prepurchase Information Seeking for New Cars and Major Household Appliances," *J. Marketing Res.*, Aug. 1972, *9*, 249–57.

Tibor Scitovsky, *The Joyless Economy*, Oxford 1976.

Harold Sheppard and A. Harvey Belitsky, *The Job Hunt*, Baltimore 1966.

E. Wolgast, "Economic Decisions in the Family," *J. Marketing*, Oct. 1958, *23*, 151 58.

Economics, Psychology, and Protective Behavior

By HOWARD KUNREUTHER AND PAUL SLOVIC*

Economics and psychology share many common interests regarding the behavior of people in the marketplace. However, these two disciplines have traditionally approached the description, prediction, and explanation of market behavior in very different ways. Psychologists have employed laboratory experiments, survey questionnaires, and some naturalistic observations to develop an empirical base of knowledge. Economists have relied heavily on utility theory and its presumption of objective rationality which, as Herbert Simon and A. C. Stedry note, "... permits strong predictions to be made about behavior without the painful necessity of observing people" (p. 272).

Over the past quarter century, a small group of economists and psychologists have been challenging the validity of the traditional economic approach. George Katona and his colleagues showed that consumer expectations, perceptions, motives, and intentions, measured by means of survey techniques, could predict economic behavior and guide public policy in situations where traditional theory was simply not adequate.

Paralleling Katona, Simon drew upon empirical research on human cognitive limitations to challenge traditional assumptions about the motivation, omniscience, and computational capacities of "economic man." As an alternative to utility maximization, Simon introduced the notion of "bounded rationality," which asserts that cognitive limitations force people to construct simplified models of the world in order to cope with it. To predict behavior "... we must understand the way in which this simplified model is constructed, and its construction will certainly be related to 'man's' psychological properties as a perceiving, thinking, and learning animal" (Simon, p. 198).

During the past twenty years, the skeleton theory of bounded rationality has been fleshed out. We have learned much about human cognitive limitations and their implications for behavior—particularly with regard to decisions made in the face of risk. Space does not permit a discussion of this work here; an extensive review is available in Slovic et al. (1976). Utility theory has been the target of repeated criticisms, both on theoretical and empirical grounds. One of the most recent and most vigorous attacks can be found in Daniel Kahneman and Amos Tversky. However, the case against the rationality of individual behavior tends to be dismissed by economists on the grounds that in the competitive world outside the laboratory, rational agents will survive at the expense of others. Thus the study of irrationality can be downplayed as the study of transient phenomena (see Simon and Stedry).

Our own experiences, as economist and psychologist collaboratively investigating people's protective actions in the face of risk, indicate that many manifestations of bounded rationality exhibited by intelligent citizens have important, nontransient social ramifications. The study of insurance behavior provides an example to which we now turn.

I. Failure of the Market in Insurance

Insurance is perhaps the oldest arrangement for shifting the financial burden from

*Kunreuther is professor and head of the department of decision sciences, The Wharton School, University of Pennsylvania. Slovic is a research associate at Decision Research, A Branch of Perceptronics, Eugene, Oregon. Support for this paper came from NSF-RANN grants ARA73-03064-A03 and ENV77-15332. Any opinions, findings, conclusions, or recommendations expressed in this publication are our own and do not necessarily reflect the views of the National Science Foundation. The insurance research described here developed out of collaboration with Bradley Borkan, Baruch Fischhoff, Ralph Ginsberg, Norman Katz, Sarah Lichtenstein, Louis Miller, and Philip Sagi.

an economic agent facing uncertain future losses to a risk-bearing institution. Economists have treated this problem as one in which the prices for different types of policies are set by the forces of supply and demand, with individuals making theoretical contingent contracts to protect themselves against different states of the world. Observed market failures have been ascribed to adverse selection and moral hazard, problems which inhibit insurers from promoting their product (see Kenneth Arrow).

Because economists have focused primarily on market mechanisms for studying social problems, they have paid relatively little attention to the impact that alternative institutional arrangements would have on behavior if an insurance market fails. Because they have assumed that individuals are utility maximizers, they have devoted little effort to studying the decision processes that individuals follow when determining whether to undertake protective action. Psychologists, on the other hand, have been actively studying risk-taking decisions by means of laboratory experiments, but only recently have begun to focus on the implications of their findings for public policy (see Slovic et al., 1976).

In this section, we will discuss a recently completed laboratory and field study that examined the decision processes involved in the purchase of flood and earthquake insurance. The details of this study are summarized in Kunreuther et al. and Slovic et al. (1977). We will describe the study and its results through the use of the simple conceptual framework depicted in Figure 1. This framework emphasizes the central importance of institutional arrangements and decision processes in developing policies for solving specific problems and should have relevance to many societal decisions involving risk.

A. *Nature of the Problem*

Natural disasters constitute an enormous problem. They annually cause several billion dollars in property damage, accompanied by an inestimable toll of human misery, anguish, and death.

The question facing public policymakers is: What are the relative costs and benefits of alternative programs for mitigating the social and economic disruption caused by natural disasters? In the case of floods, policy options that have been tried or considered include compulsory insurance, flood control systems, strict regulation of land usage, and massive public relief to victims.

B. *Institutional Arrangements*

The institutional arrangements that concern us here focused on whether or

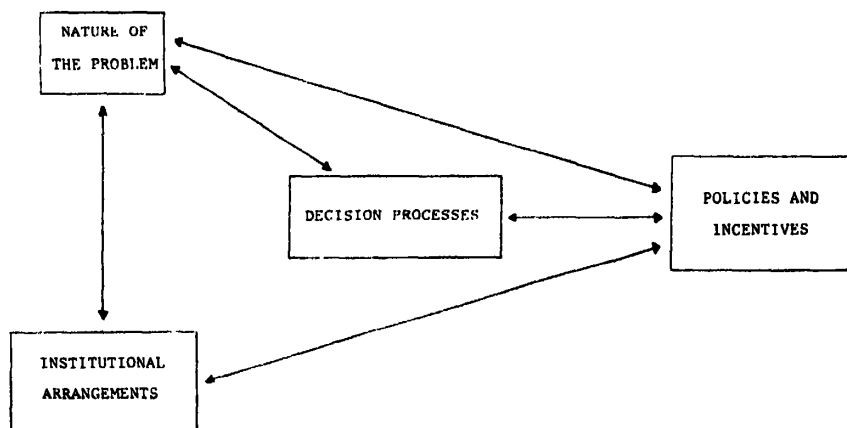



FIGURE 1. CONCEPTUAL FRAMEWORK FOR PROPOSED RESEARCH

not the purchase of disaster insurance should be required. It has been noted that, whereas few individuals insure themselves voluntarily against the consequences of natural disasters, many turn to the federal government for aid after suffering losses (see Kunreuther). As a result, the taxpayer is burdened with financing the recovery for those who could have provided for themselves by purchasing insurance. Policymakers have argued that both the government and the property owners at risk would be better off financially under a federal flood insurance program. Such a program would shift the burden of disasters from the general taxpayer to individuals living in hazard prone areas and would thus promote wiser decisions regarding the use of flood plains. For example, insurance rates could be set proportional to the magnitude of risk in order to inform residents of the hazards of living there and deter unwise development of high risk areas.

Without a better understanding of how people perceive and react to risks, however, there is no way of knowing what sort of flood insurance program would be most effective. For example, it seems reasonable that lowering the cost of insurance would encourage people to buy it. Yet, there is evidence that people do not voluntarily insure themselves against natural disasters even when the rates are highly subsidized (see Kunreuther). The reasons for failure of insurance markets need to be understood, as they have important implications for policy. Knowledge of how psychological, economic, and environmental factors influence insurance purchasing may suggest ways to increase voluntary purchases—or indicate the need for compulsory insurance programs.

C. *Decision Processes*

The primary objective of this study was to determine the critical factors influencing the voluntary purchase of insurance against the consequences of low-probability events such as floods or earthquakes. Research methods included a field survey and laboratory experiments. The field survey enabled us to discover differences between insured and uninsured homeowners in hazard prone areas, while the laboratory experiments permitted us to identify causal relationships through controlled manipulation of relevant variables.

The basic sampling plan for the field survey involved face-to-face interviews with 2,055 homeowners living in flood prone areas throughout the United States, and 1,006 homeowners in eighteen earthquake prone areas of California. Approximately half of the sample individuals were insured against flood or earthquake.

The analysis of the field survey data revealed that a significant number of homeowners in flood and earthquake prone areas either knew nothing about the availability and terms of insurance, or had inaccurate information. The survey also revealed that many residents had little idea of the probability or potential damage from a future disaster. Furthermore, the insurance decisions of persons who did have firm notions of expected losses, premium costs, etc., were often inconsistent with what would have been predicted by the expected utility model. One of the most surprising results was the large number of uninsured homeowners who expected *no* federal aid at all in the aftermath of a major disaster. This indicated that neglect of insurance could not be attributed to expectations of generous government relief.

In the laboratory experiments, subjects were presented with a series of gambles, each of which involved a specified probability of losing a given amount of money. Losses and probabilities were varied across gambles. In one experiment subjects were permitted to buy insurance against the loss at an actuarially fair rate. Additional experiments varied the premiums so that insurance was offered at subsidized rates and commercial rates. In these experiments, subjects considered well-defined insurance problems in isolation and without real stakes at risk. To supplement this format, an elaborate farm management game was designed and run by a computer. While playing this game over a five-hour period, individuals had to decide

for each year what crops they were going to plant, what fertilizers to use, and what insurance they would purchase against various natural hazards. Subjects' earnings in the game determined their salary.

The results from the experiments consistently showed that people preferred to insure against relatively high-probability, low-loss hazards and tended to reject insurance in situations where the probability of loss was low and the potential losses were high. These results suggest that people's natural predispositions run counter to some well-known economic theory (see Milton Friedman and Leonard Savage), which assumes that risk-averse individuals should desire a mechanism to protect them from rare catastrophic losses that they could not bear themselves.

When asked about their insurance decisions, subjects in both the laboratory and survey studies indicated a disinclination to worry about low-probability hazards. Such a strategy is understandable in view of the fact that limitations of people's time, energy, and attentional capacities create a "finite reservoir of concern." Unless we ignored many low-probability threats we would become so burdened that any sort of productive life would become impossible. Another insight gleaned from the experiments and the survey is that people think of insurance as an investment. Making claims and receiving payments (by insuring against more probable losses) seems to be viewed as a return on the premium. Insuring against hazards that don't occur seems a waste of money.

D. *Policies and Incentives*

Our study has led us to conclude that the primary cause of failure in the disaster insurance market is consumer disinterest. If insurance is to be marketed on a voluntary basis, then consumer's attitudes and information processing limitations must be taken into account. Policymakers and insurance providers must find ways to communicate the risks and arouse concern for the hazards. One method found to work in the laboratory experiments is to increase the perceived probability of disaster by lengthening the individual's time horizon. For example, considering the risk of experiencing a 100-year flood at least once during a 25-year period, instead of considering the risk in one year, raises the probability to .22 and may thus cast flood insurance in a more favorable light. Another step would have insurance agents play an active role in educating homeowners about the proper use of insurance as a protective mechanism and providing information about the availability of insurance, rate schedules, deductible values, etc. Of course, these actions may not be effective. It may also be necessary to institute some form of mandatory coverage, perhaps having banks and other financial institutions require disaster insurance as a condition for a mortgage.

II. Future Directions

As the world has become safer on the average, it has become potentially more dangerous at the extreme. Thus, even as technology has increased life expectancy, it has multiplied the potential for catastrophic losses due to carcinogenic chemicals, radiation releases, warfare, dam failures, etc. Reduction of technological risks typically entails substantial costs, including reduction of benefits as well. When weighing the benefits against the risks of technology, the ultimate question becomes "How safe is safe enough?" We believe that economic psychology (or psychological economics) can help provide answers to this difficult question. The framework in Figure 1 may be a useful starting point for addressing questions of acceptable risk in a way that will be helpful to system designers and policymakers. Thus, for example, when designing a set of regulatory mechanisms or incentive systems to cope with a particular hazard in society, careful attention must be paid to the current institutional arrangements and decision processes of the groups affected by that problem. Current programs imply a set of risk-benefit tradeoffs and values of life which may be inappropriate for today's society when scrutinized in this

way. On the other hand, the costs incurred by changing these current programs also have to be recognized.

At present, economists and psychologists seem to favor different approaches towards determining acceptable levels of risk. Economists have traditionally favored a market approach in which they assume that the forces of supply and demand will determine an optimal balance between the risks and benefits associated with any activity. There has been a growing recognition in recent years that environmental and technological problems involve both public and private risks. The public good (or bad) aspects of the risk call for governmental regulation or the use of other social institutions to cope with problems of market failure (see Lester Lave). These programs are typically not designed with concern for people's information processing limitations, nor are the public's judgments of risks and benefits of current and proposed systems usually considered.

Psychologists have preferred to ask people to express their risk preferences directly (see for example, Baruch Fischhoff et al.). Such an approach enables policymakers to gain insight into current attitudes and values. It also allows for widespread citizen involvement in decision making and thus has political appeal. Its principal drawback is that people may not really know what they want, or why they attach certain costs and benefits to different activities. In fact, different ways of phrasing the same question may elicit different preferences. Furthermore, people's values may change so rapidly as to make systematic planning impossible. Even if their risk preferences were stable over time, it might be difficult to translate their desires into meaningful policies without substantial implementation costs.

Policy decisions regulating risk must ultimately consider both what people say they want and what their market behavior implies they want. Thus these two approaches to assessing public preferences should be complementary rather than competing. Integrating these approaches and developing them to a level sufficient to engender public acceptance poses an exciting opportunity for collaboration between economists and psychologists.

III. Conclusion

Policymakers responsible for protecting society from natural and technological hazards need to understand the ways in which people think about risk and uncertainty. Without such understanding, well-intended policies may not achieve their goals and, indeed, may even backfire. Because rationality is "bounded," utility theory is not a trustworthy guide for policy. The understanding that is needed must come instead from systematic empirical investigations that employ multiple methods of observation and analysis.

REFERENCES

K. Arrow, "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.*, Dec. 1963, *53*, 941–73.

B. Fischhoff et al., "How Safe is Safe Enough? A Psychometric Study of Attitudes Towards Technological Risks and Benefits," *Policy Sci.*, forthcoming.

M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *J. Polit. Econ.*, Aug. 1948, *56*, 279-304.

D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, forthcoming.

George Katona, *Psychological Economics*, Amsterdam 1975.

H. Kunreuther, *Recovery From Natural Disasters: Insurance or Federal Aid?*, Washington 1973.

Howard Kunreuther et al., *Disaster Insurance Protection: Public Policy Lessons*, New York 1978.

Lester Lave, "Risk, Safety and the Role of Government," in *Perspectives On Benefit-Risk Decision Making*, Washington 1972.

Herbert A. Simon, *Models of Man: Social and Rational*, New York 1957.

——— **and A. C. Stedry.** "Psychology and Economics," in Gardner Lindzey and Eliot Aronson, eds., *The Handbook of Social Psychology*, 2d ed., Reading 1969, 269–314.

'. Slovic et al., "Cognitive Processes and Societal Risk Taking," in John S. Carroll and John W. Payne, eds., *Cognition and Social Behavior*, Potomac 1976.

——— **et al.,** "Preference for Insuring Against Probable Small Losses: Implications for the Theory and Practice of Insurance," *J. Risk Ins.*, June 1977, *44*, 237–58.

Recent Psychological Studies of Behavior under Uncertainty

By DAVID M. GRETHER*

The purpose of this paper is to survey a portion of the experimental psychology literature; viz., reports of experiments concerning choice and decision making under uncertainty. Even within this area the coverage will be far from complete—the idea being to treat only work which is both likely to be of interest to economists and to be unfamiliar to them.

The discussion which follows will be directed towards three main points. First, the behavioral regularities reported by psychologists are real, well documented, and replicatable. Thus, if it is said that "individuals in situation x exhibit behavior y," then in general one can be sure that a large number of people have been placed in the given situation and have performed as indicated. Second, much of this work seems to indicate that the behavioral assumptions employed by economists are simply wrong. For instance, choices between gambles are frequently inconsistent. In gaming situations individuals consistently do not adopt obvious optimal strategies. There are substantial and systematic biases in the perception of uncertainty. Also, individuals use information inefficiently; in particular, Bayes' rule fails as a descriptive model. For a discussion of why economists should be concerned by these results see the author and Charles Plott. Third, while these results in principle could apply to the sorts of choices dealt with in economic models, it is not as yet established that they do. In fact, in some cases close analysis of the experimental setting suggests that the results reported are precisely those predicted by conventional economic theory.

I. Probability Learning Experiments

Of all the work in experimental psychology, probably the work best known to economists is that dealing with "probability learning." Though the details of the experiments vary, the basic idea is standard. A subject is shown a sequence of Bernoulli trials, and prior to each trial the subject is asked to predict the outcome of the next trial. In spite of differences in experimental design the behavior observed is generally the same: given a sequence of trials in which the events occur with probabilities p and $1 - p$, a subject attempting to predict each trial will tend to predict the two events in proportions p and $1 - p$. That is, the relative frequency of the subject's predictions match the probabilities of the events being predicted, though the optimal strategy is to always predict the most likely outcome. (See for example Lee Roy Beach et al.)

Morris Fiorina, who surveys the prior work, notes that in most of the experiments the trials were randomized within blocks, and the observed behavior is quite reasonable once this dependence is taken into account. Thus, the probability learning experiments provide little evidence against optimizing behavior. As Fiorina says, "If anything, subjects' perceptions of the state of the experimental world are more accurate than those of the experimenters" (p. 164).

II. Bayes' Rule and Related Matters

The experiments discussed in this section deal with the way individuals process information concerning uncertainty. One of the major questions studied is whether or not subjects revise their beliefs in accord with Bayes' rule. A large number of experiments have been reported and the general finding is that people either do not revise their opinions in that fashion, or if they do, they do not use the correct "objective" probabilities in their calculations.

Many laboratory experiments of the

*California Institute of Technology.

book-bag poker-chip variety have demonstrated what has been called conservatism (see Wesley DuCharme; Beach and James Wise; Ward Edwards; C. R. Peterson, DuCharme, and Edwards). There is a tendency to treat probabilities near zero or one as being too close to one-half while probabilities near one-half tend to be correctly measured. This behavior does not appear to be replicatable outside of laboratory surroundings and many of the rules of thumb and heuristics individuals use seem to have the opposite effect (see Baruch Fischhoff; W. C. Howells, 1971, 1972; Paul Slovic, Fischhoff, and Sarah Lichtenstein).

One of the more striking examples which suggests that Bayes' rule is not a good descriptive model was reported by Daniel Kahneman and Amos Tversky (1973) and Tversky and Kahneman (1974). Eighty-five subjects were given the following instructions:

> "A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions chosen at random from the 100 available. For each description, please indicate your probability that the person described is an engineer on a scale of 0 to 100." [1973, p. 241]

Another group of eighty-six subjects was given identical instructions except that the number of lawyers was changed to thirty and the number of engineers to seventy. Both groups were given the same five descriptions to judge, and then the following: "Suppose now that you are given no information whatsoever about an individual chosen at random from the sample. The probability that this man is one of the 30 [70] engineers in the sample of 100 is __%" (1973, p. 241).

The results reported certainly do not conform to Bayes' rule. Both groups of subjects gave nearly the same posterior probabilities for each of the five descriptions in spite of the substantial change in the priors. This agreement was definitely not due to conservatism or a tendency to give probabilities equal to the prior probabilities. In fact for one of the five descriptions the median estimate of the probability that the man chosen was an engineer was around .05 for both groups of subjects and, for another of the descriptions, was around .95. In addition both groups made the "correct" response to the question quoted above. Perhaps the strangest result reported was the response to the following: "Dick is a thirty-year-old man. He is married with no children. A man of high ability and high motivation he promises to be quite successful in his field. He is well liked by his colleagues" (1973, p. 242). This description was intended to be neutral and apparently was judged so by the subjects. For both groups the median estimate was .50. Thus these subjects evaluated useless information and no information quite differently.

The hypothesis being explored in this work is that individuals make predictions based upon representativeness, and the descriptions presented were designed to test this idea. For example, one of the five descriptions was: "Jack is a forty-five-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include carpentry, sailing, and mathematical puzzles. The probability that Jack is one of the 30 [70] engineers in the sample of 100 is __%" (1973, p. 241). The results presented certainly do support the hypothesis that individuals judge by something like representativeness and ignore prior probabilities. The responses to the vacuous description and the fact that the subjects were reminded of the prior odds after each description was given are especially convincing. Nevertheless, this experiment has features that make the applicability of the findings to economic decisions doubtful. First, as is often the case, the subjects are not told the truth about the

random process being examined. Clearly, the thumbnail descriptions were not a random sample from the alleged population. The subjects' responses would agree with Bayes' rule only if they either "played the game" or believed the experimental instructions and thereby badly misperceived what was going on. Second, there is the difficulty of controlling the information given when verbal descriptions or situations are presented. Both of these difficulties could be taken care of by the use of actual balls in urns or book-bag poker-chip set ups.

For example, suppose one has two urns, one with four red balls and two white balls, and another urn with three of each color. A known randomizing device, possibly another urn or a spinner, could serve as a prior for choosing which urn to draw from. Suppose that samples of size six are drawn from one of the urns (with replacement of course). The representativeness hypothesis would seem to indicate that for samples composed of four red balls and two white balls, or three of each color, the estimated posterior odds should favor the indicated urn by more than the correct odds.

Finally there is also the question of incentives; it is not clear that Kahneman and Tversky's subjects had a positive incentive to give "correct" answers. The instructions included the following statement: "The same task has been performed by a panel of experts, who are highly accurate in assigning probability to the various descriptions. You will be paid a bonus to the extent that your estimates come close to those of the expert panel" (1973, p. 241). Thus there was an incentive to behave as the "experts" which may or may not be interpreted as attempting to give the right answer. In the ball-urn experiment, incentives can be handled by asking the subjects to guess which urn produced the sample and paying off if they guess correctly.

In fact I have run these types of experiments using economics students from several universities. The results suggest that the representativeness heuristic describes very well the behavior of financially unmotivated subjects and of financially motivated but inexperienced subjects. On the other hand the behavior of experienced subjects whose payments depend upon their decisions appears to be consistent with Bayes' theorem.

The difficulty of controlling the information conveyed in verbal presentations is shown by M. Hammerton. Ten subjects were given the following information:

"1. A device has been invented for screening a population for a disease known as psylicrapitis.
2. The device is a very good one, but not perfect.
3. If someone is a sufferer, there is a 90 percent chance that he will be recorded positively.
4. If he is *not* a sufferer, there is a 1 percent chance he will be recorded positively.
5. Roughly 1 percent of the population has the disease.
6. Mr. Smith has been tested and the result is positive. The chance that he is in fact a sufferer is: ____. . . ." [p. 252]

The median response was 85 percent with an interquartile range of 10 percent. Only one of the ten subjects underestimated the probability (which is around one half). Fourteen groups of eight subjects were given the same information except that the order of statements 3, 4, and 5 was varied and sometimes certain of them dropped. The results of those experiments are summarized in Table 1. While it is not clear just how the subjects are determining their probability estimates, it certainly is clear that they are *not* being conservative and are not using Bayes' rule.

Finally a group of twenty subjects (housewives) was given a reworded version of statements 1 through 6. Statement 1 was changed to read: "A device has been invented for screening engine parts for internal cracks" (1973, p. 253). The remaining statements were altered accordingly. For the group the median ρ was 60 percent, the interquartile range was 40 percent and seven subjects underestimated. Comparing this latter group with the population reported in Table 1 showed that the difference in the median was significant at the .001 level.

TABLE 1

Group	Presentation for Statements 3,4,5	Median ρ Estimate	Number of Underestimates
1	3,4,5	.85	1
2	3,5,4	.86	0
3	4,3,5	.85	1
4	4,5,3	.82	2
5	5,3,4	.80	1
6	5,4,3	.80	1
7	4,5	.75	2
8	3,5	.80	1
9	3,4	.75	0
10	3	.75	1
11	4	.90	2
12	5	.85	2
13		.75	1
14	3,4,5	.85	0

Source: M. Hammerton.

Kahneman and Tversky present additional evidence in favor of the representativeness hypothesis and for other heuristics also, and some quite convincing evidence that many truths of mathematical statistics are not intuitive concepts even to individuals trained in these concepts (see also Tversky and Kahneman, 1971). In particular, regression effects (for example, sampling based upon the value of the dependent variable) and sampling variability are often misunderstood. For a survey of the psychological literature see Slovic, Fischhoff, and Lichtenstein. Louis Wilde provides a survey of the evidence concerning the behavior of consumers.

III. Inconsistency in Choice

There is a substantial amount of evidence that in certain types of situations people make choices which are in some sense inconsistent. For example, though economists almost always assume transitivity of individual preference orderings, psychologists have found experimental setups that lead some individuals to choose intransitively. In a classic paper Tversky demonstrated that certain individuals will persistently demonstrate intransitivity.

Consider the following pair of gambles.

A: with probability .99 win $4.00;
with probability .01 lose $1.00.

B: with probability .33 win $16.00;
with probability .67 lose $2.00.

Note that the expected values of these gambles differ by only one cent. There is an impressive amount of experimental evidence that suggests that the behavior described below is not only possible but indeed quite common.

An individual is allowed to choose one of the two gambles and knows that he will play the gamble of his choice. In this situation he chooses gamble *A*. Instead of being asked which gamble he "likes the best," he is asked how much he would pay for gamble *A* and how much for gamble *B*. He knows that he will pay the higher of the two bids and then play the gamble he has bought. In this situation his bids will not be "unreasonable," that is, he will not bid more than $4 for *A* or more than $16 for *B*, but his bid for *B* will be higher than his bid for *A*. Alternatively, if the individual had been given the rights to play the gambles and had been interrogated as to how much he would sell them for, his responses again would be reasonable and would indicate a preference for gamble *B*. In sum, this person would choose *A* over *B*, pay more for *B* than for *A*, and would be willing to sell *A* at a lesser price than *B*.

Lichtenstein and Slovic presented 173 subjects with twelve pairs of bets similar to that shown above. In each pair one bet had a high probability of winning (*P* bet) a small amount and the other had a smaller probability of winning a large amount ($ bet). The subjects were asked for each pair which they preferred and later were asked to give selling prices. All bets were hypothetical. For 73 percent of the subjects "for every pair in which the *P* bet was chosen, the $ bet later received a higher bid" (p. 48). In a second experiment 74 subjects were asked to give their choices and buying prices. The rate of reversals was lower than in the previous case, but still significant. As before, all the bets were hypothetical. In order to verify that the phenomenon was not due to a lack of incentives, a third experiment was conducted in which the bets were played. The subjects (14) made choices and subsequently gave selling prices for the various gambles.

These basic results have been replicated, including once at a Las Vegas casino (Lichenstein and Slovic; Harold Lindman).

Economic theory suggests a number of explanations (income effects, misspecified incentives, etc.) which could explain the observed behavior. It is possible, however, to design experiments that avoid these difficulties. The author and Plott report the results of two such experiments, and to our surprise we replicated the psychologists' results. It should be noted that the gambles used were especially selected to produce the intransitivity observed, and it is not claimed that the behavior applies generally or that it would be persistent in the money pump sense.

In conclusion many of the results reported by experimental psychologists should be of interest to economists. In some cases the results obtained raise serious questions concerning the descriptive validity of a number of the behavioral assumptions used in economic models. In many instances, however, the reported behavior is consistent with economic theory, and much of this literature can be seen as showing the power of the economists' model of economic agents.

REFERENCES

L. R. Beach and J. A. Wise, "Subjective Probability Revision and Subsequent Decisions," *J. Experim. Psychol.*, Sept. 1969, *81*, 561-65.

——— **et al.,** "Probability Learning: Response Proportions and Verbal Estimates," *J. Experim. Psychol.*, Oct. 1970, *86*, 165-70.

W. M. DuCharme, "Response Bias Explanation of Conservative Human Inference," *J. Experim. Psychol.*, July 1970, *85*, 66-74.

W. Edwards, "Conservatism in Human Information Processing," in B. Kleinmuntz, ed., *Formal Representation of Human Judgement*, New York 1968.

M. P. Fiorina, "A Note on Probability Matching and Rational Choice," *Behav. Sci.*, Mar. 1971, *16*, 158-66.

B. Fischhoff, "Hindsight and Foresight: The Effect of Outcome Knowledge on Judgment under Uncertainity," *J. Experim. Psychol.: Hum. Percep. and Learning*, Aug. 1975, *1*, 288-99.

D. M. Grether and C. R. Plott, "Economic Theory of Choice and the Preference Reversal Phenomenon," *Amer. Econ. Rev.*, forthcoming.

M. Hammerton, "A Case of Radical Probability Estimation," *J. Experim. Psychol.*, Dec. 1973, *101*, 252-54.

W. C. Howells, "Uncertainty from Internal and External Sources: A Clear Case of Overconfidence," *J. Experim. Psychol.*, Aug. 1971, *89*, 240-43.

———, "Compounding Uncertainty from Internal Sources," *J. Experim. Psychol.*, Sept. 1972, *95*, 6-13.

D. Kahneman and A. Tversky, "On the Psychology of Prediction," *Psychol. Rev.*, July 1973, *80*, 237-51.

S. Lichtenstein and P. Slovic, "Reversals of Preference Between Bids and Choices in Gambling Decisions," *J. Experim. Psychol.*, July 1971, *89*, 46-55.

———, "Response-Induced Reversals of Preferences in Gambling: An Extended Replication in Las Vegas," *J. Experim. Psychol.*, Nov. 1973, *101*, 16-20.

H. R. Lindman, "Inconsistent Preferences among Gambles," *J. Experim. Psychol.*, Aug. 1971, *89*, 390-97.

C. R. Peterson, W. M. DuCharme, and W. Edwards, "Sampling Distribution and Probability Revisions," *J. Experim. Psychol.*, Feb. 1968, *76*, 236-43.

P. Slovic, B. Fischhoff, and S. Lichtenstein, "Behavioral Decision Theory," *Annual Rev. Psychol.*, 1977, *28*, 1-39.

A. Tversky, "Instransitivity of Preferences," *Psychol. Rev.*, Jan. 1969, *76*, 31-48.

——— **and D. Kahneman,** "Belief in the Law of Small Numbers," *Psychol. Bull.*, 1971, *76*, 105-110.

——— **and** ———, "Judgment under Uncertainty: Heuristics and Biases," *Science*, Sept. 27, 1974, *185*, 1124-31.

L. Wilde, "Consumer Behavior under Imperfect Information: A Survey of the Evidence," unpublished paper. Calif. Instit. Technology 1977.

DISCUSSION

GEORGE KATONA, University of Michigan: James Morgan argues convincingly that it is necessary to apply psychological principles to economics. He also gives us a useful theoretical framework by presenting five major reasons for the need for psychology in economics. He implies that even more is needed than applying psychological principles to economics. Empirical research on the behavior of consumers and business firms is required that is based on economic as well as psychological hypotheses. This is what the new discipline of behavioral or psychological economics does, of which Morgan is a successful practitioner.

I shall supplement Morgan's discussion by saying a few words 1) about the ways economic theorists exclude psychology from economic analysis, and 2) about the role of psychology in a major current issue of economic theory and economic policy, the relation of inflation to unemployment.

Recent writings on "rational expectations" indicate how some theorists make the observation and measurement of expectations unnecessary by assuming that they are formed rationally. Expectations are made an endogenous variable that may be deduced from data already included in the system (such as past trends of the variable and the consideration of past forecasting errors). This procedure is followed even though behavioral economists have measured people's expectations of incomes, prices, and some other variables for the past thirty years and have established that expectations are extrapolative only under certain specific circumstances, while under other circumstances they differ greatly from past trends.

Milton Friedman in his recent Nobel lecture refers to such psychological processes as perceptions, expectations, and surprises in order to account for the relation of inflation to unemployment. He mentions the notion of "rational expectations" with approval and states a priori what the expectations were at certain times. For instance, Friedman writes that "In the immediate post-World War II period, prior experience was widely expected to recur . . . The expectation in both countries [U.S. and U.K.] was deflation" (*Journal of Political Economy*, 1977, p. 465). But we know from extensive surveys conducted under my direction, first in Washington and then at the Survey Research Center of The University of Michigan, and published in the *Federal Reserve Bulletin*, that in 1945–46 most Americans were optimistic and expected good times to come with small price increases. In contrast to a few experts they did not expect the resumption of the deflation of the 1930's. Consumers and businessmen behaved in line with their opinions and expectations, and there was no postwar recession.

Friedman's new "tentative hypothesis," according to which "the rate of unemployment will be largely independent of the average rate of inflation" (ibid, p. 464) fails to specify the forces that make for a positive and those that make for a negative relation between unemployment and inflation. Some psychological forces operated in the first direction. I shall be brief, supplementing what I have written on the psychology of inflation in my book, *Psychological Economics*, and the book by Strumpel and myself, *A New Economic Era* (New York 1975; 1978). One of the helpful psychological principles is that about the generalization of effect: At any given time only good or only bad news is salient, and good (bad) news is thought to have only good (bad) effects. Most people believe that inflation, which is considered a bad thing, cannot have good consequences such as improving business conditions or decreasing unemployment. A second useful psychological principle says that uncertainty tends to retard, sometimes even to paralyze, action. Inflation, associated with heightened uncertainty, usually inhibits spending and investing, and stimulates saving. While stocking up and hoarding may be stimulated by the expectation of large price increases, disorientation and stress prevail, inhibiting optimism and mak-

ing for smaller rather than larger employment. Such psychological factors as the extent of uncertainty are measurable. Their strength must be determined because the net result depends on whether the negative psychological forces or the positive economic forces are stronger at a specific time.

VERNON L. SMITH, University of Arizona: I have great respect for the scientific contributions of Paul Slovic and his associates. Their experiments establish that the majority of people in uncertain situations tend to violate such criteria as "simple" dominance in utility theory. These results have been replicated in a wide variety of contexts, some with significant rewards. Although there remains much to be learned about the dynamic persistence of these phenomena under "money pump" and other learning conditions, these results should and must be taken seriously by economists.

However, I disagree with much of the Kunreuther-Slovic interpretation of these results. Consider utility theory, which in changing form has been around for over two centuries, which has been enormously important in the development of a well-defined concept of rational choice, and without which the referenced experiments could not even have been designed. Some theory of utilitarian choice is likely to be around long after the present generation is buried. To be viable, such a theory must take into account the fact that thinking, transacting, information processing, information interpretation, and learning are costly acts for people. What the cited experiments do is to reject the validity of traditional utility theories that ignore these costs. It was in this spirit nearly twenty years ago that the psychologist, Sidney Siegel, in *Annals of New York Academy of Sciences, 1961*, using the binary choice paradigm, demonstrated the inadequacy of both nonutilitarian psychological theories of choice and traditional utility theory for explaining his experimental results. He demonstrated the explanatory power of expanding utility theory to include costly decisions.

Instead of damning utility theory, I think Howard Kunreuther and Paul Slovic would be well advised to begin the more useful business of modifying it in the light of their empirical results. A theory of rationality is self-contradictory if it does not incorporate the subjective costs of decision. Economists will certainly not object to such modification. Utility concepts are much too useful to be discarded so casually. I would also suggest that it is naive to suppose that supply and demand theory in particular, and the theory of decentralized markets in general, are incapable of dealing with subjective decision costs. Kunreuther and Slovic have been reading obsolete textbooks and believing what they read.

I also have several objections to the Kunreuther-Slovic interpretations of economics and of the alleged policy implications of their experimental results.

1. Economics does not argue that in the "competitive world . . . rational agents will survive at the expense of others." Kunreuther and Slovic are using "competitive" in its strict biological sense. Economics argues that under competition (i.e., where there is free choice among alternatives) people have an incentive to specialize in those activities in which they have a comparative advantage and then to cooperate through market exchange to capture the gains from trade. Communication between economics and psychology is not served when scholars in either discipline so inaccurately summarize the position of the other.

2. Kunreuther and Slovic at several points write as if insurance against hazard was an objective good, for the rational and irrational alike; that people too ignorant to know what is "obviously" good for them need to be either manipulated; or that the power of the state should be used to coerce both the "right" decision and the means of financing it. Perhaps it is indeed the case that *1984* is only six years away.

3. The fact that a large number of uninsured homeowners expected no federal aid in the aftermath of a major disaster may also suggest that such homeowners are willing to assume the risk of such low probability events.

4. It is entirely consistent with utility

theory that limitations on time, energy, and "attentional capacity" reduce the search for insurance information. Utility theory does not predict that if *A* is preferred to *B*, then an individual will switch from *B* to *A* unless the expected cost of making the change is less than the expected benefit.

5. If people view insurance as an investment by making claims and receiving payments from insurance against probable losses, why is this irrational? If major hazards involve risk of life as well as property loss, why not invest in insurance against those events for which one is likely to be around to collect the claim.

6. Instead of mandatory flood insurance, I would suggest a simple label to be attached to property deeds in flood plain areas: *Warning, geologists have determined that flood plains sometimes flood; and psychologists have determined that if you are ignorant enough to buy property here, you may be too ignorant to realize that you should buy flood insurance.*

Sex Differences in Labor Supply Elasticity: The Implications of Sectoral Shifts in Demand

By CYNTHIA B. LLOYD AND BETH NIEMI*

The recent recession has been distinguished both by the persistence of inflation throughout the downturn and by the failure of the recovery to lower the unemployment rate to its prerecession level. It has been suggested that the continued rapid growth in female labor force participation throughout the recession has been affected by this inflation, and has also been one possible cause of the high postrecession unemployment rates.

The question this paper addresses is whether the elasticity of labor supply with respect to employment conditions, both over the business cycle and in the long run, has changed over time, and, if so, which demographic groups have contributed to the change. A brief summary of the relevant labor supply theory will be followed by an examination of recent patterns of labor force entries and withdrawals. Then we will present our research strategy and empirical results, which embody two alternative approaches to the measurement of business cycle conditions and long-term trends. Finally, the implications with respect to both future changes and policy needs will be considered.

The research reported here documents the existence of both short-run and long-run shifts in the labor supply elasticities of men and the importance of the industrial composition of the demand for labor in determining both sex differences and changes in labor supply elasticity over the business cycle. However, such shifts in demand do not provide an explanation for observed long-run changes, for both men and women, in the elasticity of labor supply with respect to the expected real wage.

The life cycle theory of labor supply predicts that individuals and family units plan their total lifetime commitment to the labor force on the basis of permanent income expectations, but time their participation to take advantage of short-run fluctuations in economic conditions. In the long run, preferences for leisure versus income interact with changes in wage rates and real income to determine secular changes in labor force participation. In the short run, the substitutability of nonmarket and market work for each individual, and the substitutability of family members in nonmarket and market work determine the responsiveness of labor supply to temporary fluctuations in economic opportunities.

The observed labor supply response to temporary change in employment opportunities is the net outcome of the familiar discouraged and additional worker effects. The empirical studies in this area indicate a net procyclical behavior of the labor force. Results from cross-section and time-series studies differ in their estimates of the degree of labor force sensitivity, but generally agree that the responsiveness of women and of older and younger men to the business cycle is greater than that of prime-age men. In part because of such findings, women and younger and older men are often referred to as "secondary" workers and prime-age men as "primary" workers. The major explanation of these sex and age differences in labor supply elasticity is the differential importance of nonmarket

*Assistant professor of economics, Barnard College, and associate professor of economics, Newark College of Arts and Sciences, Rutgers University, respectively.

activities for different groups. Because the attractiveness of these alternative opportunities does not fluctuate over the business cycle, the relative advantage of market work is directly related to the business cycle.

I. Recent Changes in Labor Force Flows

An examination of labor force entry and withdrawal by men and women over the recent recessions of 1970–71 and 1974–75, as presented in Table 1, suggests some changing patterns. In each recession, the number of labor force entrants of both sexes declined. The decline in the number of male entrants was much greater in 1974–75 than in 1970–71, as might be expected given the severity of the more recent recession. However, it is somewhat surprising that the decline in female entrants in 1974–75 was smaller than in 1970–71. Increased withdrawals were another indicator of the discouraged worker effect among men, but the number of withdrawals among women actually declined during each recession, indicating that the decrease in labor force quits among the employed outweighed the increase in the exit of the unemployed from the labor force.

Labor force growth increased with the economic recovery for both men and women. In 1971–72, entries increased and withdrawals declined for both sexes, but there was no increase in the rate of entry for women in 1975–76. Consequently, their increased labor force growth stemmed entirely from a decline in withdrawals. These summary statistics suggest that women may have changed their patterns of labor force entry and withdrawal over the business cycle. However, it is difficult to determine from this evidence alone to what extent the observed changes are related to short-run fluctuations versus longer term trends.

TABLE 1—TOTAL CIVILIAN LABOR FORCE CHANGE (1000's): 1967–76

	Net Change	Left[a]	Entered[b]
Males			
1967–68	+ 546	3424	3970
1968–69	+ 688	3669	4357
1969–70	+ 974	3660	4634
1970–71	+ 826	3707	4533
1971–72	+1244	3562	4806
1972–73	+ 938	3714	4652
1973–74	+ 983	3776	4759
1974–75	+ 429	3894	4323
1975–76	+ 744	3723	4467
Females			
1967–68	+ 844	6329	7173
1968–69	+1308	6507	7815
1969–70	+1008	6470	7478
1970–71	+ 571	6392	6963
1971–72	+1186	6062	7248
1972–73	+1233	6329	7562
1973–74	+1315	6495	7810
1974–75	+1173	6218	7392
1975–76	+1416	5961	7377

Source: Employment and Earnings.

[a] Estimated from persons out of the labor force who left a job within the previous 12 months. For example, to estimate the number who left between mid-1967 and mid-1968, an average was taken of the four quarterly figures for 1968. Because these four quarterly figures include persons who left jobs from the beginning of 1967 to the end of 1968, an average gives a more accurate estimate of the flow between two mid-years. Obviously, certain assumptions must be made to convert a stock figure into an estimated flow.

[b] The number of those who entered is calculated as the sum of the net change and the number who left.

II. Research Strategy

Our research strategy involves estimating labor supply elasticities with respect to unemployment and real wages from quarterly data for two time periods (1956–65 and 1966–76) and testing for statistically significant shifts in these elasticities between the two periods. Regressions were run for three aggregate labor force groups (total, male and female 16+) and six age-sex groups (men and women aged 16–24, 25–55, and 55+). All equations were run in *log*-linear form, using the Cochrane-Orcutt iterative technique.

The first equation relates civilian force participation rates to a linear trend, three seasonal dummies, and the aggregate unemployment rate, lagged one quarter. In the second equation, labor supply elas-

ticities are estimated with respect to the expected real wage, defined as the product of the employment rate $(1-UN)$ and the real wage (*REAL*), lagged one quarter. A measure of structural shifts in the sex composition of the demand for labor (*EMPW*) is also included as an independent variable in this formulation, as well as the three seasonal dummies.

There are strong theoretical as well as practical reasons for including the real wage. In all cross-section studies, the effects of income and wage changes are seen as the primary determinants of labor supply in both the short and long run. Unfortunately, in the time-series context, it is not possible to separate income changes from wage changes, or husbands' wage changes from wives' wage changes, so that the effect of the real wage on labor force participation must be interpreted as the net outcome of income, substitution, and cross-substitution effects. Also, because of the unusual importance of inflation in the 1970's, the real wage must be included in order to test accurately for a shift in labor supply behavior over time. By not entering the money wage rate and the price level separately, we implicitly assume the absence of any significant money illusion. (See Ray Fair and/or Michael Wachter for more detailed discussion of this point.)

The inclusion of a variable measuring sex-specific shifts in the composition of the demand for labor was suggested by William Bowen and T. Aldrich Finegan. Because of sex differences in occupational and industrial distributions, secular and cyclical changes in the composition of demand will have different impacts on men and women (see Ralph E. Smith). The variable *EMPW* is interpreted as measuring the share of women in total employment if women's share in each industry's employment remains fixed and the overall industrial structure varies over time.

Because both *REAL* and *EMPW* exhibit strong secular trends, as well as cyclical fluctuations, it is not possible in this second equation to distinguish between short-run and long-run labor supply responses. Given the upward trends in both these independent variables, the linear trend term is no longer necessary and, in fact, its inclusion only creates instability in the regression coefficients due to high multicollinearity.

III. Empirical Results

Table 2 presents the labor supply elasticities with respect to both the probability of employment and the expected real wage, as estimated from equations (1) and (2). In equation (1), the only statistically significant net discouragement effects in 1956–65 were for young men and older women, and no net discouragement is observed for the labor force as a whole. However, the cyclical sensitivity of the labor force appears to have increased over time. Male labor supply elasticity is statistically significant in 1966–76 and the F-test shows a significant difference between the two periods. Although both young and older men show significant discouragement effects in the second period, the difference in supply elasticities is most significant among prime-age males. Young and prime-age women also appear to show a discouragement effect in the second period, although the difference in coefficients between the two periods does not prove to be statistically significant. Older women show a change from net discouragement to a net added worker effect which, although not a statistically significant change, might result from the effects of husbands' earlier retirement on wives' labor market behavior.

A comparison of the employment elasticities estimated for the two equations reveals that the coefficient on the employment rates changes systematically when additional explanatory variables are added. The employment coefficient in equation (2) is significantly positive for almost all groups in the first period. In the second period, the employment coefficient is less positive for men and more positive for women.

The inclusion of *EMPW* in equation (2) may explain part of the difference in the employment coefficients obtained in the two equations. Because *EMPW* is much

TABLE 2—LABOR SUPPLY ELASTICITIES WITH RESPECT TO PROBABILITY OF EMPLOYMENT AND EXPECTED WAGE

	Total 16+	Males 16+	Females 16+	Males 16–24	Females 16–24	Males 25–54	Females 25–54	Males 55+	Females 55+
Employment Elasticities									
Equation (1)[a]									
1956–65	.091	.017	.220	.366[e]	.469	−.024	−.059	.114	.677[d]
1966–76	.198[f]	.257[f]	.228	.615[d]	.733[f]	.064[e]	.354[d]	.332[d]	−1.30
Direction of Change	+	+	+	+	+	+	+	+	−
F-Ratio	1.90	7.80[f]	.19	.20	.33	6.13[e]	.64	.95	1.23
Equation (2)[b]									
1956–65	.331[f]	.175[d]	.688[e]	.463[e]	.642	.023	.460[d]	.645[e]	1.226[e]
1966–76	.295[e]	.047	.808[f]	−.783	.831[d]	.013	.866[f]	.493[d]	.996
Direction of Change	−	−	+	−	+	−	+	−	−
F-Ratio	.03	.23	1.00	2.47	1.05	.03	.60	−	−
Real Wage Elasticities[c]									
Equation (2)									
1956–65	.088[f]	−.252[d]	.821	−.015[d]	.596	−.037	.785[d]	−.455[e]	1.507
1966–76	.355	.040	.661	.055	947	−.056	.681[d]	.678	−.384
Direction of Change	+	+	−	+	+	−	−	+	−
F-Ratio	−	3.92[e]	.65	4.56[e]	.72	.10	.21	.56	1.79

Sources: Employment and Earnings and unpublished Bureau of Labor Statistics data.

[a] $\ln LFP = a' + b_1 UN + b_2 T$ where LFP = the civilian labor force participation rate of the demographic group in question, UN = aggregate unemployment rate lagged one quarter $\doteq -\ln(1 - UN)$; T = a linear trend (1 . . . 84).

[b] $\ln LFP = a' + b_1 UN + b_3 \ln REAL + b_4 \ln EMPW$ where $REAL$ = the real wage, defined as "gross hourly earnings of production or nonsupervisory workers or private nonagricultural payrolls" divided by the consumer price index, lagged one quarter

$$EMPW = \sum_{j=1}^{11} a_j E_{jt} \Bigg/ \sum_{j=1}^{11} E_{jt}$$

where E_{jt} = total employment in industry j in quarter t, and a_j = the average proportion female of employment in industry j over the 21-year period.

[c] Real wage elasticity calculated as $b_3 - b_1$

[d] Significant at 10 percent confidence level

[e] Significant at 5 percent confidence level

[f] Significant at 1 percent confidence level

(Significance levels for real wage elasticities are determined by the least significant of the two coefficients.)

more highly correlated with the unemployment rate in the second period than in the first period, it has a systematic effect on patterns of coefficient change. This correlation suggests that recessions in the 1970's had a greater negative impact on the employment in male-dominated sectors than in the first period. The result is that, whereas *EMPW* is only significant and positive for women in the first period, it is significant and negative for men as well in the second period.

It is clear from these results that the employment elasticities in equation (1), which are usually interpreted as supply responses to short-run changes in employment conditions, are in fact a joint measure of supply responses to fluctuations in employment rates and sectoral shifts in demand. Therefore, what appeared to be a statistically significant shift in supply elasticities in equation (1) was in fact partially due to a sectoral shift in demand, unfavorable to men and favorable to women.

In addition to this demand effect there appear to be shifts in labor supply behavior in response to the expected wage, shifts which probably represent long-run changes. In the first period, the expected wage elasticities are consistently negative for men and positive for women. These results conform to traditional patterns, with income and cross-substitution effects dominant for men and the substitution effect dominant for women. In the second period, these patterns change substantially. The direction of change is not inconsistent with the aggregate results in the first equation, with men's labor supply behavior becoming more procyclical or less countercyclical, and women's labor supply becoming less procyclical. However, the changes are statistically significant only in the case of total male and young men's labor supply, whose responsiveness to change in the real wage decreased dramatically over the period.

These results suggest that discouragement is not solely a function of employment rates but rather of all the demand factors which affect employment conditions over the business cycle. Although the second equation yields more accurate measures of the unemployment coefficient, neither equation provides an entirely correct measure of discouragement, because of the impossibility of disentangling short-run from long-run change.

IV. Implications

Given the extreme disparity between men and women in the occupational distribution, it would be naive to assume that one measure of employment conditions could accurately reflect shifts in job opportunities for both men and women. In order to compare discouragement effects it would be necessary to estimate sex-specific employment probabilities which would clearly be dependent on sectoral patterns of demand. Without such evidence, we must be cautious in our designation of secondary versus primary worker status, since our results suggest that women's procyclical sensitivity has decreased while men's has increased. While the regression results show no significant change in labor supply behavior for women between the two ten-year periods chosen, the labor force flow data in Table 1 suggest the possibility of such a change in the 1970's. It would seem that a distinction between primary and secondary workers would be more accurately based on age than sex. Over the long run, it appears that women's labor supply relationship is becoming less positive and men's is becoming less backward bending.

The relatively favorable demand conditions facing women appear to reinforce their long-run flow into the labor force and to create a less discouraging economic environment during recessions. However, in the long run, rising female labor force participation cannot be accomodated solely by economic growth which simply reinforces present patterns of demand. Despite favorable growth in industries traditionally employing women, the gap between women's actual share in total employment and what it would be if their representation within industries remained constant (as measured by *EMPW*) has widened over time. Since policies and institutions clearly interact with and reinforce one another, it is to some extent misleading to discuss general policy implications in the absence of any reference to the existing legal and institutional framework. The distribution of the benefits of prosperity and growth will depend in large part on whether past discriminatory patterns are allowed to perpetuate themselves. On the other hand, the opening up of new opportunities to women will be much more feasible in a context of sustained economic growth than under conditions of continuing economic stagnation. Improvements in the functioning of labor markets, from both the supply side in terms of providing opportunities for training and mobility and from the demand side in terms of eliminating various types of discrimination, can be of great importance in the achievement of the general economic goal of prosperity and noninflationary full employment.

REFERENCES

William Bowen and T. Aldrich Finegan, *The Economics of Labor Force Participation*, Princeton 1969.

R. C. Fair, "Labor Force Participation, Wage Rates, and Money Illusion," *Rev. Econ. Statist.*, May 1971, *53*, 164-68.

J. Mincer, "Labor Force Participation and Unemployment: A Review of Recent Evidence," in Robert A. Gordon and Margaret S. Gordon, eds., *Prosperity and Unemployment*, New York 1966.

R. E. Smith, "The Impact of Macroeconomic Conditions on Employment Opportunities for Women," study prepared for the use of the Joint Economic Committee, Congress of the United States, *Achieving the Goals of the Employment Act of 1946—Thirtieth Anniversary Review*, *1*, employment paper no. 6, Washington 1977.

M. L. Wachter, "A Labor Supply Model for Secondary Workers," *Rev. Econ. Statist.*, May 1972, *54*, 141-50.

U.S. Bureau of Labor Statistics, *Employment and Earnings*, various issues.

Women's Increasing Unemployment: A Cross-Sectional Analysis

By R. CHRISTOPHER LINGLE AND ETHEL B. JONES*

Disaggregation of unemployment rates by sex reveals two relationships. The more frequently recognized is that female rates have tended to exceed male rates since World War II. Another aspect of recent concern is the apparent worsening of this difference during the 1960's (see *Economic Report of the President*; George Perry). In this paper we use cross-section data from the Censuses of Population of 1960 and 1970 to examine whether the relationship between female and male unemployment rates shifted during the decade. We first present a model for examining the structure of female unemployment rates using cross-section data and then test for a parameter shift between 1960 and 1970 in the relationship between male and female unemployment. The data are generated for the ninety-nine Standard Metropolitan Statistical Areas (*SMSA*) of 250,000 or more population in both 1960 and 1970.

I. The Model

The model specified relates the unemployment rate of women to the unemployment rate of men, while recognizing that female rates also may differ among geographic areas because of particular characteristics of the female labor force and labor markets. The specification for estimation is:

$$(1) \quad OFUN = \alpha_0 + \alpha_1 PMUN + \alpha_2 MARF + \alpha_3 F > 45 + \alpha_4 MERN + \alpha_5 EDC + \alpha_6 RACE + \alpha_7 FIND + \alpha_8 FERN + \mu$$

where

OFUN = civilian unemployment rate of women 20 years of age and over

PMUN = civilian unemployment rate of males 25–54 years of age

MARF = percent of the total female labor force married and living with spouse

$F > 45$ = percent 45 years of age and older of the female labor force of 20 years of age and over

MERN = for 1960, the median annual income of men who worked fifty to fifty-two weeks in 1959 and, for 1970, the median annual earnings of men who worked fifty to fifty-two weeks in 1969

EDC = median education of the population of women 25 years of age and over

RACE = percent nonwhite of the total female labor force

FIND = an index (see William G. Bowen and T. Aldrich Finegan; Judith Fields) of the extent to which the area's labor market is oriented toward employment of women

FERN = the counterpart data for women of *MERN*

μ = a residual

The expected values of α_1, α_2, and α_6 are greater than zero; the expected values of α_3, α_5, and α_7 are less than zero; and, the values of α_4 and α_8 are ambiguous, a priori. The dependent variable is measured for women beyond the teenage years because labor studies usually treat teenage unemployment as a separate problem.

Discussion of the independent variables follows the approach of recent years that unemployment is a function of turnover and the duration of search (see Charles C. Holt). The search process is an income-maximizing strategy in which the worker weighs the returns and costs of search (see George Stigler). For women, turnover must consider the spells of unemployment both

*Assistant professor and professor of economics, respectively, Auburn University. We wish to acknowledge the helpful comments of Richard Higgins.

between jobs and between labor force entry and exit for full-time home activity. The potential importance of labor force turnover in the unemployment experience of a woman requires consideration of a vector of variables identifying her relationship to the household (marital status, age, and husband's earnings) as well as vectors denoting her market characteristics (education and race) and attributes of the labor market for women of the community in which she resides (the level of female earnings and the extent to which the market affords job opportunities to women).

A. *Male Unemployment*

Male unemployment is that of males 25–54 years of age. This age range removes differences among areas in the impact upon unemployment of variation in the concentration of students and of variation in the accelerated rates of decline during the 1960's of the labor force participation of older men. The a priori expectation of $\alpha_1 > 0$ obtains simply from the observation that male and female unemployment rates have generally moved together during the post-World War II period (see Beth Niemi, p. 332). If the relationship between male and female unemployment shifted over the decade, the parameter estimate of α_1 is expected to change between 1960 and 1970.

The ratio of female to male unemployment varies over the cycle, increasing during expansion and falling at times of contraction (see Niemi, p. 331). Using the specific investment hypothesis, Niemi observed that the expected smaller amounts of specific investment in women should lead to higher layoff rates for women than men during cyclical downturn, but that the effect of the higher layoff rates was dampened by the cyclical timing of women's labor force participation and the smaller cyclical sensitivity of the occupations and industries in which women are concentrated.

While economic activity was slow during the early 1960's, the late 1960's was a period of economic expansion. The observed male unemployment rate (*PMUN*) may be viewed as having two components: the usual level based upon the particular labor market conditions of a city (U_F) and a cyclical component (U_C) with the coefficients γ_F and γ_C, respectively. The declining relationship of female to male unemployment as total unemployment rises implies that $\gamma_C < \gamma_F$. If U_C is larger in one year (t_1) than in another (t_2), α_1 would be larger in t_2 than in t_1 by virtue of the cyclical component. The average male unemployment rate across the *SMSA*'s was 4.06 in 1960 and 2.63 in 1970. Hence, we expect α_1 to be larger in 1970 than in 1960 because of the implied larger cyclical component in *PMUN* in 1960.

B. *Household Characteristics*

The marital status variable controls for the expected higher unemployment level of wives that accompanies their more frequent reentrance into the labor force because changes in home responsibilities produce interruptions in their attachment to the labor market. The spells of unemployment would be more frequent since reentrance may include a period of unemployment. The duration of a spell of unemployment may also be longer for married women if the availability of another income source (the husband's income) encourages more selectivity in search (see Nancy S. Barrett and Richard Morgenstern, p. 458). Thus we expect $\alpha_2 > 0$.

The age variable is a proxy for the relationship between the family life cycle stage and the unemployment experience of married women. Older married women would experience fewer interruptions in labor market activity associated with child care, and, hence, fewer of the spells of unemployment that accompany labor force reentry. Among single women, those 45 years of age and older have fewer years of labor market activity remaining over which the gains of search can be realized. This factor should reduce their spells of unemployment to search for another job, implying $\alpha_3 < 0$.

The variable *MERN* is a proxy for the earnings level of the husband. Lower levels

of *MERN* are expected to increase the continuity of the married women's labor market attachment in order to add to the family's income stream. This continuity should reduce the spells of unemployment associated with labor force reentry but will increase the spells of unemployment accompanying job termination since she is less likely to leave the labor force. The duration of search, and hence the level of unemployment, may increase at higher levels of *MERN* if higher levels of *MERN* enable the woman to be more selective and less pressured in her job search process. Thus α_4 is of ambiguous sign.

C. *Education and Race*

Niemi has presented data that show the level of female unemployment falls with increasing levels of education. Insofar as amounts of formal schooling and specific on-the-job investment are positively correlated, the increased wage associated with specific investment (see Gary Becker) will raise the worth of market time relative to home time and encourage continued labor force activity. The continued labor force activity reduces the spells of unemployment from labor force turnover. The spells of unemployment from voluntary job leaving to search for another job are also less frequent since the increased wage from larger amounts of specific investment reduces the gains from search. The average duration of a spell of unemployment may also be negatively related to education level because women with more education may have "... more knowledge of labor markets and more labor market information ..." (Niemi, p. 339).

Several studies have attributed the higher rates of unemployment of blacks to more frequent spells of unemployment (see Barrett and Morgenstern; Robert Hall). Discrimination, by blocking the job ladders to higher paying employment opportunities, places blacks in employment opportunities of relatively less specific investment. The number of spells of unemployment are increased because job termination by the employer costs less for workers of less specific investment. Also, since jobs of less specific investment involve less wage loss from job movement, black workers would have a higher propensity to leave current employment and search for another job.

D. *Market Characteristics*

The 1973 *Economic Report of the President* stressed increasing labor force participation of women as a major factor in the growing disparity of the 1960's. However, Barbara Bergmann and Irma Adelman emphasized the market concentration of women in particular industries and occupations as the cause of the growing disparity. The duration of search would increase as relatively more women competed for "women's" jobs. The variable *FIND* provides a measure of the percent of an *SMSA*'s employment that is oriented towards women. Increases in *FIND* should decrease the duration of search to locate a potential job opening and hence be negatively associated with the female unemployment rate.

The relationship between the level of unemployment and the proxy for the wage level of women across *SMSA*'s (*FERN*) is ambiguous because of the opposing pulls upon both duration and spells. A higher wage level would be expected to reduce the duration of a search because of the implied increased opportunity cost of search time. However, the number of spells may increase because the higher wage level raises the relationship of market to home productivity, thus inducing labor force entry and market search. Insofar as increasing job opportunities do not accompany the increasing number of entrants, the spells of unemployment will rise.

II. Empirical Results

Because of correlation among the residuals (0.55) of the ordinary least squares estimates of the model for 1960 and 1970, Arnold Zellner's seemingly unrelated regression technique (*SUR*) has been used. This technique has the advantage of providing *F*-statistics for testing for equality

TABLE 1—SEEMINGLY UNRELATED REGRESSION EQUATIONS OF FEMALE UNEMPLOYMENT, 1960 AND 1970

	1960 (1)	1970 (2)	F-Statistic (3)	1960 (4)	1970 (5)	F-Statistic (6)
Constant	4.82	8.17	.72	3.03	7.25	.99
	(1.26)	(2.43)		(0.80)	(2.02)	
PMUN	.56	.88	17.37	1.03	1.12	.007
	(8.41)	(12.89)		(4.78)	(3.57)	
$(PMUN)^2$	-	-	-	-.04	-.03	.01
				(2.27)	(0.71)	
MARF	.05	.05	.01	.05	.05	.02
	(1.70)	(1.94)		(1.65)	(1.93)	
F >45	-.004	-.04	1.73	-.008	-.04	1.31
	(0.12)	(2.26)		(0.26)	(2.18)	
MERN	-.0007	-.0002	1.64	-.0005	-.0002	.77
	(1.67)	(1.25)		(1.39)	(1.27)	
EDC	-.36	-.14	1.13	-.36	-.12	1.27
	(2.43)	(0.70)		(2.54)	(0.58)	
RACE	.01	.01)	.01	.008	.01	.31
	(0.97)	(1.42)		(0.62)	(1.49)	
FIND	-.12	-.14	.14	-.09	-.14	.63
	(2.01)	(3.16)		(1.52)	(3.02)	
FERN	.002	.0003	14.96	.002	.0004	10.88
	(4.12)	(1.29)		(3.85)	(1.30)	

Note: *t*-ratios in parentheses.

between parameter estimates of the two years. The *SUR* results are reported in Table 1 (columns (1) and (2)) together with the *F*-statistics (column (3)).

All parameter estimates for which the signs were predicted a priori perform as predicted. The signs of the two variables for which the relationship involved ambiguity are consistent between 1960 and 1970. A higher level of female earnings is accompanied by increasing levels of female unemployment, and the female unemployment rate falls as the proxy for husband's earnings increases. At the 5 percent level of significance the *F*-statistics indicate a shift in the size of the coefficients for only *PMUN* and *FERN*.

While the *F*-statistics suggest that particular coefficients are equal, for a number of the variables the *t*-statistics are considerably different, as in the case of education. For this reason, identification of the significant variables has been made by observing the *t*-values after pooling the observations for both 1960 and 1970 and using the information from the *F*-statistics. Regression equations were estimated in which the coefficients of all variables except *PMUN* and *FERN* were restricted to be equal in both years. By this method, the variables *FIND*, *MARF*, and *FERN* are significantly different from zero at better than the 5 percent level, and the education variable is significant at the 10 percent level.

The findings on the "femininity" index lend support to the labor market segmentation argument of Bergmann and Adelman. *Ceteris paribus*, job searching by women is improved at higher values of *FIND*, most likely because a job offer can be located more quickly if the structure of employment in the area is more oriented towards industries that employ women. The sign and significance level of *MARF* also suggest increasing participation as a factor in the growing disparity, since married women represented an increasing component of the labor force during the decade as a result of their increasing participation rates. Increases in the earnings of women and their level of education influence women's unemployment, but in

offsetting directions. Racial composition, male earnings, and age are not important variables, *ceteris paribus*, in the determination of variations in the level of female unemployment.

The difference in the values of the coefficient of *PMUN* and the *F*-test statistic imply a structural change in the relationship between men's and women's unemployment. However, the earlier discussion of the model held that estimates of the coefficient of *PMUN* would be affected by the level of business activity. The cyclical relationship suggests that female unemployment rises at a slower rate at higher levels of male unemployment. We test for this relationship by introducing the square of *PMUN* into the model. The *SUR* regressions, inclusive of $(PMUN)^2$, together with the accompanying *F*-statistics, are reported in columns (4)–(6) of Table 1.

When $(PMUN)^2$ is in the model, the coefficients of *PMUN* in 1960 and in 1970 have similar values. The sign of the coefficient of $(PMUN)^2$ is in accord with the observation of a decline in the rate of increase of female unemployment as male unemployment increases. The *F*-statistic no longer suggests a structural shift for the coefficient of *PMUN*, and the coefficient of *FERN* is the only one not equal between the two equations at the 5 percent level. Again the *t*-values of the coefficients of particular variables are of differing orders of magnitude between the two equations although the *F*-statistics are small. When the equations are reestimated subject to the restraint of equality of the coefficients in both years except for *FERN*, *PMUN* and $(PMUN)^2$ have *t*-values indicating significance at the 1 percent level. The findings with respect to significance levels for the other independent variables remain as reported earlier.

III. Conclusion

This study adds the cyclical factor to the list of changes during the decade of the 1960's that would influence any trend in the disparity between male and female unemployment rates. The cyclical effect, taken in isolation, is not a small one. Between the dates of the highest (1961, 5.7 percent) and the lowest levels (1969, 2.1 percent) of male unemployment during the decade, the unemployment rate disparity of men and women 20 years of age and over increased by 1.0 percentage points. If the highest and lowest values of the male rates were substituted into the 1970 equation inclusive of $(PMUN)^2$ (all coefficients except *FERN* restricted to be equal between 1960 and 1970), the difference between male and female unemployment would have increased by 0.9 percentage points due to the changed level of male unemployment at the two dates.

While accounting for the cyclical factor may simply imply no structural change in the relationship between male and female unemployment during the decade, the cyclical characteristic of women's unemployment behavior does have an uncomfortable aspect. Recent research has emphasized the discontinuity of women's labor market attachment in accounting for the differences in wages between men and women (see Jacob Mincer and Solomon Polachek). The cyclical aspect of the unemployment behavior arises from women leaving the labor force. While the cyclical aspect, on the one hand, reduces the amplitude of cyclical swings in unemployment, at the same time it is masking the problem of the effect of economic fluctuations upon the continuity of women's labor market experiences.

REFERENCES

N. S. Barrett and R. D. Morgenstern, "Why Do Blacks and Women Have High Unemployment Rates," *J. Hum. Resources*, Fall 1974, 9, 452-64.

Gary S. Becker, *Human Capital*, 2nd ed., New York 1975.

B. R. Bergmann and I. Adelman, "The 1973 Report of the President's Council of

Economic Advisers: The Economic Role of Women," *Amer. Econ. Rev.*, Sept. 1973, *63*, 509–14.

William G. Bowen and T. Aldrich Finegan, *The Economics of Labor Force Participation*, Princeton 1969.

J. Fields, "A Comparison of Intercity Differences in the Labor Force Participation Rates of Married Women in 1970 with 1940, 1950, and 1960," *J. Hum. Resources*, Fall 1976, *11*, 568–77.

R. E. Hall, "Turnover in the Labor Force," *Brookings Papers*, Washington 1972, *3*, 709–56.

C. C. Holt, "Job Search, Phillips' Wage Relation, and Union Influence: Theory and Evidence," in Edmund S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 53–123.

J. Mincer and S. Polachek, "Family Investments in Human Capital: Earnings of Women," in Theodore W. Schultz, ed., *Economics of the Family*, Chicago 1973, 397–429.

B. Niemi, "The Female-Male Differential in Unemployment Rates," *Ind. Labor Relat. Rev.*, Apr. 1974, *27*, 331–50.

G. L. Perry, "Changing Labor Markets and Inflation," *Brookings Papers*, Washington 1970, *3*, 411–41.

G. S. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–25.

A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *J. Amer. Statist. Assn.*, May 1962, *57*, 348–68.

U.S. Bureau of the Census, *U.S. Census of Population: 1960*, Vol. 1, *Characteristics of the Population*, Part 1 and appropriate state parts, Washington 1963; 1964.

———, *U.S. Census of Population: 1970*, Vol. 1, *Characteristics of the Population* and appropriate state parts, Washington 1973.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1973.

Unemployment Rate Targets and Anti-inflation Policy as More Women Enter the Workforce

By CLAIR VICKERY, BARBARA R. BERGMANN, AND KATHERINE SWARTZ*

As women's labor force participation rates have continued to increase, it has become commonplace to argue that the targets that policy planners set for average unemployment rates should be adjusted upwards to "correct" for the fact that women's unemployment rates have historically been higher than men's.[1] In this paper, we argue that most "target corrections" are based on an oversimplified approach to labor market realities. This approach has tended to promote the attitude that the high unemployment rates of women are an incurable and unregrettable fact of nature, and has also tended to bias policy discussions in a direction that leads to the toleration of overall slack and puts low or zero emphasis on the labor market problems of women.

Abstracting from issues of age and race, we may characterize the usual target correction methodology as starting with the estimation of a simple relationship between women's and men's unemployment rates (U_w and U_m) such as

$$(1) \quad U_w = a + b\,U_m$$

and then going on to calculate a "corrected" target unemployment rate U^* by

$$(2) \quad U^* = p(a + b\,U_m^*) + (1 - p)\,U_m^*$$

where p represents the current proportion of women in the labor force, and U_m^* is the level achieved by the men's unemployment rate the last time we were at "full employment."

Our objection to such calculations derives from our unwillingness to accept the implicit assumption that equation (1) is a basic fact of nature. We view (1) as the result of a set of calculations based on numerical data generated in the historical past and influenced strongly by past macroeconomic policy, social choices, failure to enforce antidiscrimination laws, and other human choices which may well be different in the future. We consider it valuable to look beyond such formulations as (1) to representations that portray more realistically the situation of women in the labor market. Such efforts should allow us to speculate in a more informed way about the interplay of future macro and "structural" policies and their potential effects.

There are two major factors which cause women to have higher unemployment rates than men do: women have a higher rate of turnover into and out of jobs and into and out of the labor force than do men; and in a labor market divided by occupational segregation of the sexes into men's "turf," and women's "turf," the supply-demand balance in the women's turf is more unfavorable than in the men's. In exploring the importance of these two factors at some length, we shall argue the following:

1. Under current conditions of occupational segregation by sex, women's higher turnover rate contributes little to the excess of women's unemployment rates over men's rates in recessionary periods.

*Vickery is assistant professor of economics, University of California-Berkeley. Bergmann and Swartz are, respectively, professor and assistant professor of economics at the University of Maryland. Bergmann acknowledges support from the Computer Science Center of the University of Maryland and that some of the materials incorporated in this work were developed with the financial support of the National Science Foundation Grant 77-14693. However, any opinions, findings, conclusions, or recommendations expressed herein are our own and do not necessarily reflect the views of the Foundation.

[1]One result of such thinking was Aaron Gordon's calculation of unemployment rates "corrected" for changes in the age-sex-race composition of the labor force.

Even if women had had the same turnover rates as men but the degree of occupational segregation had remained the same, neither women's unemployment rates nor average unemployment rates would have been much affected in the slack labor markets of the mid-1970's.

2. Differences in voluntary turnover between women and men become more important as a source of different unemployment rates between men and women as the economy approaches full employment, but the effect of these differences on an appropriately computed target "full-employment" unemployment rate are slight.

3. The increased female participation has occurred primarily because each successive female cohort is displaying an increased attachment to the labor force. This trend among the younger cohorts should result in a decline in the labor turnover among women on average. Nevertheless it will also bring on a deterioration in the position of women in the labor market unless occupational segregation is ended.

4. Occupational segregation is currently the major culprit in raising women's unemployment rates above those of men. However, the rigor of occupational segregation is affected by the degree of slackness in the economy. The persistence of a high level of demand might be expected to reduce occupational segregation by sex *below its historical levels*. Moreover, policies to open up jobs to women will have a better chance of success in an economy where there is not a sizeable excess supply of men.

I. Turnover, Group Unemployment Rates, and Average Unemployment Rates

In this section we demonstrate that women's higher turnover rates, under present conditions of sex segregation and slack in the labor market, contribute relatively little to women's unemployment rates and little to overall unemployment rates. We also show that if sex segregation of the labor market were diminished, women's higher turnover rates would become more important in creating a differential between men's and women's unemployment rates, but paradoxically, the contribution of women's high turnover rates to average unemployment rates would still be minor.

The relationship between turnover, sex segregation, and unemployment rates can be seen most clearly by first assuming a situation in which the size of the labor force and the number of job slots are constant, allowing us to abstract from the influence of changing demand and supply conditions. These assumptions imply that a separation of a person from a job creates a net addition to vacancies and unemployment simultaneously. Whether the labor market is segmented or unsegmented by sex, the impact of an additional separation from a job on the total number of unemployed people will depend upon the length of time the job opening remains unfilled. If every separation were immediately followed by an accession, the rate of unemployment would be entirely unrelated to the separation rate. If, on the other hand, there were a delay of M months between a separation and the accession which filled the resulting vacancy, then each extra separation per month would add M to the number of unemployed persons. Whatever the value of M (Bergmann has estimated that it is on the order of .3),[2] the addition to the number of unemployed persons caused by additional separations is equal to the number of additional vacancies so caused. (Under our assumption of constant size labor force, knowing the rate at which women enter and leave the labor force will not provide additional information about their unemployment rate beyond that provided by knowing the rate of their separation from jobs.)

But what is the impact of a turnover on group unemployment rates? Maintaining our assumption that no job slots are created

[2]This estimate of the average length of completed periods of vacancy was made by running the simulation model described in Bergmann, in which the parameters were fitted to track the unemployment rate, unemployment durations, and vacancy data of 1969–73.

or destroyed and that the number (although not the personal identity) of those in the labor force remains constant, let us consider two extreme cases:

CASE A: *There is perfect sex segregation of jobs and job hunters*. Consider the effect of an extra female separation on women's unemployment rates if women are off by themselves in a walled-in corner of the labor market. That extra separation cuts both ways: it has the effect of creating a newly unemployed woman, but it also has the almost offsetting effect of opening up a vacancy into which some other unemployed woman will very shortly go. The women's unemployment rate will be affected very little if M is small, and men's unemployment rates will be affected not at all.

CASE B: *Gender identity is irrelevant in the labor market*. If both sexes are competing for whatever vacancies open up, then an extra separation by a women will open up a vacancy which very shortly may be filled by a man. In this case, the extra woman's separation is likely to increase women's unemployment rate to a far greater extent than was true in the first case. Furthermore, the extra woman's separation also has the effect of decreasing the unemployment rate of men. However *average* unemployment rates will be affected only by vacancies.

More formally, under our assumptions unemployment rates in the perfect segregation case can be represented by

$$(3) \quad U_i = [LF_i - D_i (1 - v_i)]/LF_i \qquad i = \text{male, female}$$

where LF_i is the size of the women's (or men's) labor force, D_i is the number of persons employers would like to have in female (or male) occupations, and v_i is the vacancy rate for female (or male) jobs. In this case, the difference between women's and men's unemployment rates will primarily reflect different supply-demand imbalances in the men's and women's markets. In the perfect integration case, the ratio of women's to men's unemployment rates will, in the steady state, be equal to the ratio of the women's average flow into unemployment from job separations to the men's average flow, so that

$$(4) \quad \frac{U_w}{U_m} = \frac{s_w (1 - U_w)}{s_m (1 - U_m)}$$

where the s_i are the separation rate for women or men. In this case, the separation rates are the primary determinant of relative unemployment rates. In either case, the average unemployment rate U will then be

$$U = [LF_m + LF_w - D(1 - v)]/(LF_m + LF_w)$$

where D represents total job slots and v the average vacancy rate. Labor turnover becomes an important determinant of the unemployment rate as the number of job slots approach the size of the labor force and as the duration of vacancies increase.

While it is true that increased female participation brings on an increase in the average turnover rate, the extent of the damage to the unemployment rate may be judged from the results of a simulation done earlier by Vickery. She assumed women's mobility patterns remained unchanged but that occupational segregation was considerably reduced. Using the 1976 labor force composition and allowing an increase in demand which would bring the adult white male unemployment rate down to 2.3 percent resulted in a total unemployment rate of 3.9 percent, which is only slightly higher than the 3.8 percent witnessed in 1966.[3]

II. The Age-Specific Characteristics of the Female Labor Force

Although women of all ages have been increasingly prone to participate in the

[3]The adult white male unemployment rate fell considerably further, to 1.9 percent in 1968 and 1969. For the simulated adult unemployment rates, see Vickery. In the above target rate, the unemployment rate for teenagers was assumed to be the same as in 1966. If occupational segregation were not assumed to change, the target unemployment rate would have risen to 4.2 percent.

labor force, the influx of women into paid work has occurred largely through the introduction of successive female cohorts characterized by considerably greater participation rates than earlier cohorts have shown. While total female participation rates increased 24 percent from 1961 to 1976, the participation rate of women under 30 increased 42 percent and the rate of women 30 and over increased 13 percent. The increased market work of younger women has been particularly potent in its impact on the labor force because it has coincided with the entry of the "postwar baby boom" children into the labor market, a demographic occurrence that began in the mid-1960's and will continue until 1980.

The participation patterns of younger women indicate that women's attachment to the labor force is growing. To the extent that this is true, we should expect the women's average rates of labor turnover to decline, which provides a double-edged outcome—frictional unemployment among women will decline, but women workers will no longer provide a buffer stock of workers to cushion the swing in the unemployment rate.

III. Prospects for the Desegregation of Occupations

With the continued heavy inflow of women into the labor market, improving the opportunities for women's employment becomes an increasingly important part of an overall strategy for reducing the economic waste and personal loss associated with high unemployment rates. While some of the increasing numbers of women who want to work are being absorbed by increases in employment in the traditional "women's ghetto" occupations, placing heavy reliance on that solution would, in the unlikely event it succeeded, invariably bring a deterioration in women's wage position relative to men. Women coming into the labor market must be accommodated increasingly by opening up jobs previously closed to them in the professional, technical, managerial, administrative, and crafts occupations.

Some progress in occupational desegregation is already occurring. In the five years since 1972, increased employment of women in professional, technical, managerial, administrative, and crafts jobs accounted for 57 percent of the 3.6 million increase in total employment of women. In this time the proportion of women among professional and technical workers has risen from .39 to .43 and among managers and administrators from .18 to .22. The rapidly rising enrollments during the past decade of women in business, medical, and law schools, and other professional degree programs indicate that the pool of workers eligible for these jobs is increasingly populated by women. As a consequence, we should expect to see the percentage of women in these occupations progressing toward 50 percent. On the other hand, the percentages of women in blue collar jobs are not changing appreciably. They remain very low—particularly for craftsmen and foremen, where since 1972 the proportion of women has gone from .04 to .05.

It is important to note that the progress that has been made in the desegregation of occupations has taken place in the face of slack demand, and despite the fact that the government agencies charged by law with enforcing nondiscriminatory employment practices are judged to have been of very low efficacy by the Civil Rights Commission and most other observers. Market forces—the increasing availability of women and the increasing gap between men's and women's pay—have been working in this direction and the changing ideas concerning the place of women in the labor market have also no doubt had some effect. If in the future a better enforcement effort is mounted, along with a monetary and fiscal policy which has the effect of eliminating slack more rapidly, it would not be unreasonable to expect quite rapid progress in occupational desegregation.

IV. Policy Implications and Prescriptions

We have argued here that higher turnover rates of women cannot be used to justify upward revisions of unemployment

targets now that the rate of women participating in the labor market has increased. Instead, it is our contention that occupational segregation is the primary cause of higher women's than men's unemployment rates, and that occupational segregation can be attacked successfully.

Progress against occupational segregation will mean a reduction of the inflationary pressure associated with a given level of the average unemployment rate. In past periods, when unemployment rates have diminished, tightness has developed first in those parts of the labor market which have been the traditional preserve of white males. Policies which serve to increase the number of workers whom employers are willing and able to hire for traditionally white male jobs will spread around the slack in the labor market more evenly, and thus will permit lower levels of unemployment corresponding to a given degree of tightness in supply. A policy which successfully works towards the reduction in occupational segregation will leave more "room" for the action of policy instruments in raising the demand for goods and services. The action of the latter will in turn make occupational desegregation easier.

Programs designed to counter supply-side problems in the labor market must also be geared to reducing women's unemployment rates. The low degree of improvement in women's employment in crafts jobs—13 percent of all jobs currently—reflects the fact that women have been passed over as candidates in the more desirable employment and training programs, in spite of the fact that 41 percent of families below the poverty level in 1976 were single parent or single person households, where the parent or single person was an able bodied woman of working age.

An unemployment policy which consists of a judicious package of measures—to reduce slack, reduce occupational segregation and reduce the number of people without marketable skills—will improve the welfare of the currently disadvantaged groups, while continuing to provide white males with low unemployment rates. If carried on with sufficient skill and vigor, such a policy might well lead to lowering average unemployment rate targets rather than raising them.

REFERENCES

B. R. Bergmann, "Empirical Work on the Labor Market: Is There Any Alternative to Regression Running?," *Proc. 27th Annual Meeting Ind. Relat. Res. Assoc.*, Univ. Wisconsin-Madison 1974, 243–51.

Robert Aaron Gordon, *Disaggregating the Goal of Full Employment*, Washington 1977.

C. Vickery, "The Impact of Turnover on Group Unemployment Rates," *Rev. Econ. Statist.*, Nov. 1977, *59*, 415–26.

DISCUSSION

BARRY CHISWICK, Hoover Institution, Stanford University: R. Christopher Lingle and Ethel Jones are concerned with the recent increase in the ratio of adult female to adult male unemployment rates. Using data on ninety-nine *SMSA*s from the 1960 and 1970 *Censuses of Population*, and a set of quite reasonable explanatory variables, they estimate equations to explain the unemployment rate of adult women.

They find no structural change from 1960 to 1970 in the cross-section female unemployment rate equations, once they allow for the non-linear effect of male unemployment rates. Thus some of the rise in the ratio of female to male unemployment rates is due to lower unemployment rates for both sexes in 1970, and the smaller cyclical sensitivity of the female rate. Unfortunately, they estimate the effect of the change from 1960 to 1970 of only one explanatory variable, the male unemployment rate, and changes in other variables may have been important.

Lingle and Jones also find that the greater the relative importance of the "male intensive" industries the higher the female unemployment rate. They do not note, however, that male intensive industries have declined in relative importance as a source of male employment in recent decades. Thus I would suspect that the unemployment rate of men across *SMSA*s may be related positively to the relative importance of male intensive industries. Moreover there is an important, but often misunderstood, difference between the concepts of the male intensiveness of an area's industrial structure and the extent of sex segregation of employment in the area.

Perhaps the most important secular factor in female unemployment is the rise in the relative importance of married women in the female labor force. Is the effect of marital status attributable to the greater movement of married women in and out of the labor force? This hypothesis could have been pursued by adding labor force attachment variables. The 1960 and 1970 Census micro data tapes include information on labor force status in the reference week (the variable used by Lingle and Jones), weeks worked in the previous year, and whether the person worked five years ago. Thus variables could be constructed for the proportion of adult women in the labor force in the reference week in each *SMSA* that were employed last year and five years ago.

Cynthia Lloyd and Beth Niemi are concerned with "whether the elasticity of labor supply with respect to employment conditions . . . has changed over time." They estimate the elasticity of age-sex specific civilian labor force participation rates (*LFPR*) with respect to the overall unemployment rate (*UN*), real average hourly earnings (*REAL*), and the relative importance of female-intensive industries in the economy (*EMPW*). They use quarterly data for two subperiods, 1956–65 and 1966–76.

They conclude that there has been a narrowing of sex differences in procyclical labor force participation. That is, women's procyclical sensitivity to the probability of employment and wages has not changed, while for men, particularly young men, it has increased. However they offer no theoretical explanation for their findings.

The labor force status questions in the *CPS* were revised in 1967 in a way that may have significantly altered sex differences in unemployment, and hence in the labor force participation rate. It is not clear whether the Lloyd-Niemi findings are a result of this change in the survey instrument. Their hypothesis could have been tested by including interaction terms for a continuous time variable with both *UN* and *REAL* in each subperiod or, in an equation computed for the entire period, the interactions of a continuous time variable and a post-1967 dummy variable with *UN* and *REAL*.

It is now well established that the overall unemployment rate has undergone a secular drift and is an inferior cyclical index. It also introduces a simultaneous equation bias, which is particularly severe for adult women. For example, an exogenous decline in fertility raises the attachment of women to the labor market, alters the fe-

male unemployment rate, and hence changes the overall unemployment rate. The "prime-age" male unemployment rate is a better cyclical index.

Clair Vickery, Barbara Bergmann, and Katherine Swartz assert that occupational segregation is "the major culprit in raising women's unemployment rates above those of men" and they belittle the impact of the greater interlabor force mobility of women. These assertions apparently conflict with the data.

Women are disproportionately employed in service and white collar jobs in which the unemployment rate is relatively low for both sexes. If occupational segregation causes greater unemployment among women by decreasing the probability of finding a job, we should have observed that the increased female participation is associated primarily with a longer duration of their unemployment. Yet the higher female rate appears to be due to a greater incidence of unemployment.

In the last low unemployment year, 1973, the unemployment rate of adult women exceeded that of adult men by 1.6 percentage points. When unemployment rates are computed excluding unemployed new entrants and reentrants to the labor force from the data, the women's rate exceeds the men's rate by only 0.2. If we also exclude unemployed voluntary job leavers from the labor force data, the resulting unemployment rate (for job losers) for women is less than the rate for men by 0.2. (See *Economic Report of the President, 1975*, p. 103.)

It would seem that the occupational distribution of women in the economy tends to lower their relative unemployment rate, while their greater interlabor force mobility tends to raise it.

MARIANNE A. FERBER, University of Illinois: The topic for this session, "The Effects of Increased Labor Force Participation of Women on Macroeconomic Goals," is of considerable interest to both scholars and policymakers. Particularly crucial (to those who have not defined unemployment out of existence) is the question whether the continued influx of women into the labor market requires an adjustment in the definition of full employment. Equally important is the question of whether a higher female proportion in the labor force makes it more difficult to achieve both an acceptably high rate of employment and low rate of inflation at the same time.

The Clair Vickery, Barbara Bergmann, and Katherine Swartz paper squarely addresses both these issues. They argue persuasively that the higher turnover rate of women raises unemployment only moderately in a tight labor market and hardly at all in a slack labor market. However they recognize that a level of demand sufficiently high to permit the employment of continuously increasing numbers of women in "female" occupations would cause inflationary pressures in "male" occupations. The breakdown of occupational segregation is the solution they suggest.

This analysis of the problem and the prescription for a solution both appear sound. At the same time their reading of recent developments is rather optimistic. The proportion of women among professional and technical workers has increased, but substantial segregation within that category continues. The rising enrollments of women in business, medical, and law schools is impressive, but progress toward 50 percent in an occupation is found to be far slower than progress toward 50 percent among entrants.

While neither of the other papers focuses directly on the effect of the influx of women on macroeconomic goals, each is relevant to the relation of the female labor supply to the unemployment rate.

The Cynthia Lloyd and Beth Niemi paper is primarily concerned with sex and age differences in the elasticity of labor supply, and focuses largely on the importance of the sex composition of the demand for labor. They conclude that there has been a shift in demand for labor over the business cycle and secularly which is favorable to women and unfavorable to men. The tendency for male employment to vary procyclically in more recent years is largely

explained by this change, as is the decline in procyclical variation of female employment.

There is no reason for questioning the facts. Industries traditionally employing women have expanded relative to others and tend to be less prone to cyclical fluctuations. It is not clear, however, whether the causal relationship is as one-sided as implied in this paper. Could it be that the increasing tendency for women to remain in the labor force results in a (relative) decline in their wages during a recession while employment remains relatively high? This subject remains to be investigated. Could it be that the continuing influx of women into the labor market contributes to the expansion of female occupations? The decline of (relative) wages in these occupations tends to support this hypothesis, as does the secular increase in female relative to male unemployment.

The R. Christopher Lingle and Ethel Jones paper investigates what they call the apparent worsening of female compared to male unemployment and concludes there has been no structural change in the relationship. Four troublesome questions arise. First, they conclude there has been no significant structural change because none of the parameters have changed significantly. There may, however, be a significant change in the constant. Hence, while the conclusion is justified that the slope remained the same, there may have been an upward shift nonetheless. Second, a previous study by myself and Helen Lowry in *Signs: Journal of Women in Culture and Society*, Spring 1976, found that change in military forces significantly affects the ratio of female to male unemployment. As it happens there was a decrease of about 250,000 in military forces in 1970 (while the decrease in 1960 was only about 25,000) which would cause male unemployment to be unusually high that year. Third, attributing higher unemployment to the increased participation of married women as opposed to single women is difficult to substantiate empirically since the two are highly correlated. Fourth, while the parameter for male unemployment did not increase significantly, it did increase. It is altogether plausible that had the change been tested over a longer period of time, say since 1950, or 1945, it would have been significant.

Thus, the evidence that there has been no change in the relationship between female and male unemployment has not increased relative to that of males is not fully convincing. Nor is it clear that it is primarily favorable developments on the demand side that have made possible the absorption of increasing numbers of women in the labor market. Hence Vickery, Bergmann, and Swartz's emphasis on appropriate policies to reduce occupational segregation and maintain overall high employment rates is entirely warranted.

RALPH E. SMITH, The Urban Institute: The fundamental issue suggested by the title of this session is whether the feminization of the labor force *should* lead to a reduction in our macroeconomic goals regarding unemployment. The three papers presented here each deal with specific aspects of women's labor market activities in recent years, but only the paper by Clair Vickery, Barbara Bergmann, and Katherine Swartz (V-B-S) addresses the unemployment goal issue. Rather than commenting on the theoretical and measurement problems in each of these papers, I will focus my remarks on the general theme of this session.

A logical starting point is to enumerate the potential mechanisms by which higher aggregate unemployment could result from women's increased labor force activity. Then research can address methods of dealing with each. I can think of four contenders.

First, women usually have had a higher unemployment rate than men. The simple standardization adjustments criticized by V-B-S implicitly assume that for a given level of aggregate demand this difference is immutable. This, of course, is unlikely. For example, if women's labor force attachment increases and if occupation segregation is reduced, the difference would narrow.

Second, even if the female-male unemployment gap were eliminated, the sheer size of the increase in the total labor supply resulting from women's increased participation could lead to higher unemployment for both women and men. Fiscal and monetary policies may not have been sufficiently stimulative to absorb the additional labor supply. This is a more plausible link between higher aggregate unemployment and the behavior of women. However, this argument deals with unemployment achieved, not the goal. Since the economy's potential grows more rapidly as more women come into and remain in the labor force, this is an argument for increasing the employment target, not the unemployment target.

Third, macroeconomic planners are probably basing their aggregate demand policies on inadequate labor supply forecasts. If potential *GNP* is growing more rapidly than they anticipate, then the planners, not the women, should be blamed.

The first three hypotheses suggest that increased female labor force size could make it *more difficult* to achieve lower aggregate unemployment. My fourth hypothesis is that policymakers consider lower unemployment a *less desirable* goal when there are more women in the labor force. Sexism is not necessary to support this hypothesis. One of the benefits of being in a multi-earner family is that family income is less affected by an individual's job loss. The increase in participation by wives acts as an income stabilizer, possibly reducing the hardship associated with the unemployment of either the husband or the wife. On the other hand, more wives in the labor force may make families more dependent on the labor market and less capable of filling their needs within the nonmarket economy. Furthermore, many more women are now in the labor force because they must support themselves. Thus policymakers may be in error if they are assuming that higher unemployment is now less serious.

Although sexism on the part of policymakers is not necessary to support this hypothesis, sexism may well exist. Policymakers may view women's labor market activity as less important than that of men and consequently view their lack of employment less seriously. I think an extremely fruitful area of research is the cost of female unemployment. Is there any justification for viewing women's unemployment as less serious than men's?

I believe that all four hypotheses help explain why women's increasing labor forces participation has resulted in higher unemployment in recent years. More research than has been reported here will be required to assess the relative importance of these factors and, more importantly, to forecast the outlook for the future.

But it should be stressed that, even if the female-male unemployment gap were eliminated, three of the explanations of the increased difficulty in achieving low unemployment would remain. Sharp increases in female participation, especially if unanticipated, may make it harder for macroeconomic policy to work, and may make policymakers less anxious to try very hard.

Planning for a Resource-Rich Region: The Case of Alaska

By DAVID T. KRESGE and DANIEL A. SEIVER*

Development of the nation's natural resources, particularly its energy resources, has become a matter of increasing concern in recent years. One cause for concern is that such developments are frequently of large scale and can have major impacts on the region in which the resource is located. We report on a model which has been developed to estimate the regional economic impacts of resource development and, more specifically, evaluate regional policies designed to deal with these impacts. The model is used to analyze the situation confronting Alaska as its petroleum resources are developed to meet the nation's energy needs. The results obtained from the Alaska model are of direct interest because Alaska is such a prominent part of the overall *U.S.* energy picture. In addition, Alaska offers an excellent laboratory for a general analysis of the resource development process. Although the magnitudes of the development projects in Alaska are unusually large, this does not change the nature of the process; it merely makes the impacts easier to identify.

We give below a brief description of the structure of the Alaska model, along with a few summary statistics from historical situations. The model is then used to examine the implications of several major fiscal policy strategies available to the state. Finally, some tentative guidelines are offered for the design of effective policy strategies in regions experiencing major resource developments.

*Associate director, Harvard-M.I.T. Joint Center for Urban Studies, and faculty research fellow, National Bureau of Economic Research, Inc.; and assistant professor of economics, University of Alaska (on leave). This work was supported by the National Science Foundation. Assistance from Edward Porter, Scott Goldsmith, and Michael Scott is gratefully acknowledged.

I. An Alaska Model[1]

We have divided the regional economy into "export" and "residentiary" sectors. Production levels in the export sectors are specified exogenously, since output is constrained either by the availability of natural resources or by federal policy decisions.[2] Outputs in the residentiary industries are determined by Alaska incomes, prices, and other local demand conditions. Employment in each industry is calculated from a labor requirements function, or inverse production function. The Alaskan price level is determined jointly by *U.S.* consumer prices (almost all Alaska consumer goods are imported) and by local demand conditions. Alaska personal income consists chiefly of wages and salaries (sector wage rates are functions of *U.S.* wages and local conditions). Subtracting federal and state income taxes, as determined by the fiscal model, and deflating by the Alaska price level produces an estimate of real disposable personal income, which is the key variable determining the outputs of the nonexport industries.

In terms of its general structure, our model is similar to the regional model archetype presented by Norman Glickman. The population and fiscal submodels, however, go well beyond most regional model specifications. In the population

[1]A complete description of the model structure is given in Kresge et al. The data sources are described in Kresge (1974a, b).

[2]The principal export sectors are: petroleum; agriculture, forestry, and fisheries; fish processing; wood and paper products; and the federal government.

model, for example, the amount of civilian net migration to the state is determined endogenously as a function of Alaska employment growth and income in Alaska relative to the United States as a whole.[3] Thus rapid employment growth attracts substantial migrants, particularly if the new jobs pay high wages (for example, oil pipeline construction). This relationship is quite strong over the historical period, and has been reconfirmed by the experience during the pipeline construction period 1974–76. Migration flows keep Alaskan incomes from diverging excessively from *U.S.* levels.[4] The demographic model incorporates an age-sex-race distribution of the state's population, which when combined with a set of (exogenous) age-sex-race-specific fertility and mortality rates, and a standard aging process determines the natural increase of the population.

The Alaska model has a detailed fiscal sector in which each major source of revenue is estimated as a function of the tax-structure and the relevant measure of economic activity.[5] Government expenditures are modeled by functional category, each of which has an associated employment intensity. The level of total state government expenditures is a key policy variable which is specified by the policy alternative being analyzed. Local government revenues are also modeled by major source. State revenue sharing, a key component of local revenues, is explicitly incorporated. Total local government expenditures are assumed to be equal to revenues.

The seventy-five stochastic equations in the model have been estimated using ordinary least squares (*OLS*) with annual data

TABLE 1—MEAN ABSOLUTE PERCENT ERROR (*MAPE*) STATISTICS—ALASKA MODEL

Variable	*MAPE*
State revenues	1.05
Gross output	2.35
Total employment	2.49
Personal income	2.82
Population	3.63

for the period 1961–74.[6] Using the true values of exogenous and policy variables, the model was simulated over the 1964–74 period, which was characterized by relatively steady economic growth culminating in a full-scale boom resulting from oil pipeline construction. The mean absolute percent errors (*MAPE*) of the key endogenous aggregates for this period are presented in Table 1. Although there are no official standards to judge the model's historical accuracy, these *MAPE* statistics are not greatly different from those reported for other regional models.[7] This model has been used to generate the policy simulations which are discussed below.

II. Fiscal Policy Strategies

There is little question that the future growth of the Alaska economy will be dictated largely by the rate of development of Alaska's petroleum resources.[8] Within this general setting, however, the state has a wide range of policy options that it can use to influence the pattern and, to some extent, the pace of economic expansion. Perhaps the single most important decision confronting the state of Alaska concerns the extent to which revenues derived from nonrenewable resource development are

[3]The precise income variable is the lagged ratio of real per capita disposable personal income in Alaska to the equivalent *U.S.* measure. Net migration is allocated to age-sex groups based on 1970 Census patterns. For additional detail, see Seiver.

[4]Alaska real wages can remain above *U.S.* real wages indefinitely, given migration costs, imperfect and costly information, and certain Alaska climatic disamenities.

[5]For additional detail, see Goldsmith.

[6]Data availability occasionally reduced the sample period to 1964–74. Two-stage least squares with principal components (*TSLSPC*) was also tried, producing results quite similar to *OLS*.

[7]See Glickman, pp. 164–65.

[8]Our basic petroleum development scenario (including a constant real price of oil) relies heavily on the Federal Energy Administration. The effects of alternative assumptions, such as a falling real price of oil, and more rapid petroleum development are discussed in Kresge et al.

TABLE 2—SELECTED SIMULATION RESULTS: ALASKA MODEL

		Population (000's)	Employment (000's)	Real Disposable Personal Income (1967 $ billions)	State Expenditures ($ billions)	State Accumulated Surplus ($ billions)
	1975	405	199	1.24	0.78	0.38
Case 1: 25 Percent	1980	487	239	1.66	2.19	.87
Savings Rate	1985	576	277	2.06	4.23	2.18
	1990	633	300	2.41	4.57	3.75
Case 2: 50 Percent	1980	479	234	1.63	2.04	1.38
Savings Rate nondeclining	1985	564	271	2.02	4.01	4.00
real per capita expenditures	1990	659	321	2.59	5.58	4.85
Case 3: 6 Percent	1980	455	218	1.51	1.51	2.53
Growth in real per	1985	495	232	1.72	2.54	12.32
capita expenditures	1990	619	310	2.49	5.22	21.42
Case 4: 8 Percent	1980	460	221	1.54	1.61	2.35
Growth in real per	1985	520	248	1.84	3.15	10.26
capita expenditures	1990	718	373	3.00	8.08	10.19
Case 5: 50 Percent	1980	487	239	1.69	2.08	1.16
Reduction in	1985	582	280	2.11	4.17	2.73
personal income tax	1990	690	336	2.74	5.91	1.29

set aside for use when the resource is exhausted. Alaskans have already voted to divert a minimum of 25 percent of oil royalties into a permanent fund, the principal of which cannot be spent on current account.

When the Alaska model is used to simulate the effect of a 25 percent savings policy, the economy is projected to grow throughout the entire period to 1990 but the growth is very unsteady (Case 1). Employment, for example, increases at 5.2 percent per year from 1978–85, but only 1.6 percent per year for 1985–90.[9] By holding to a fixed savings rate, the state discards the possibility of using fiscal policy to stabilize Alaska's long-run growth path. The level of operations in the petroleum industry would be the primary determinant of economic activity in Alaska. In fact, this type of passive state fiscal policy, by causing state spending to move in tandem with petroleum activity, will actually accentuate the fluctuations caused by varying rates of petroleum development.

A fixed savings policy, in addition to increasing the magnitude of economic fluctuations, becomes untenable by the late 1980's. State revenues level off and then begin declining after 1985 as production from Prudhoe Bay passes its peak. With revenues declining and with a fixed savings rate, state spending falls, and real state spending per capita drops by 15 percent between 1985 and 1990. The implied drop in the level of government services being provided to Alaska residents is likely to make this policy politically infeasible, as well as economically imprudent.

To avoid some of the problems associated with a fixed savings rate, the state could set aside more in earlier years in order to have funds available to sustain spending in later years. To simulate the effects of this type of policy, it is assumed that 50 percent of royalties are saved during those years when it is possible to do so without having to cut back on real state spending per capita (Case 2). In later years, when petroleum revenues begin to decline, the savings rate is gradually reduced in

[9]Selected simulation results are presented in Table 2. Users of large scale simulation models are often buried by reams of "output" and some economists consider this a fitting punishment. Others may wish to see more, and are referred to Kresge et al.

order to keep real state spending per capita from declining.

With this policy, higher levels of state spending are supported in the later years by the earnings on the permanent fund, which reaches more than $5 billion compared to $3 billion with a 25 percent savings rate. By saving a larger proportion of petroleum earnings in the early years, the state converts a temporary surge in revenues into a continuing income stream. This has the effect of moderating the employment growth rate in the 1978–85 period and producing much stronger growth of 3.5 percent a year from the period 1985–90. This is clearly a much less erratic growth path than that achieved under a 25 percent savings rate.

While the 50 percent savings rate damps some of the more extreme fluctuations, it does not lead to the accumulation of sufficient balances to completely eliminate the problems caused by "boom-bust" resource developments. By 1985, the savings rate has begun to fall below 50 percent in order to sustain public services, and by the end of the period, the state will be confronted with the uncomfortable prospect of having to cut back on public services, raise taxes, or draw down the balances accumulated in the "permanent" fund. Thus the savings-expenditure policy considered here cannot be maintained indefinitely, though it is much more viable than the fixed 25 percent savings rate.

Under either of the fiscal policy options considered thus far, spending patterns in the years immediately after the completion of the trans-Alaska oil pipeline set the stage for the state's eventual fiscal difficulties. Because of the tremendous surge of state petroleum revenues (from $0.5 billion in 1977 to $1.4 billion in 1980, and $2.8 billion in 1983), and even with a 50 percent savings rate, real per capita state expenditures are projected to more than double between 1977 and 1983. With production from Prudhoe Bay, and hence state revenues, expected to start declining in the mid-1980's, the increase in spending is simply not sustainable in the long-run. Perhaps even more importantly, the surge in state spending accelerates the pace of economic activity and attracts more people into the state. The model projects an increase of 75,000 in Alaska's population between 1977 and 1980, two-thirds of which is induced interstate migration.

A larger population combined with a higher level of per capita spending could drive state expenditures so far above current revenues that the entire "permanent" fund could be exhausted in just a few years. One way of dealing with this problem would be to undertake severe austerity programs in the late 1980's by curtailing public services, cutting government employment, and raising taxes. An alternative approach would be to deal with the initial source of the problem, namely, the surge in state spending in the late 1970's. If state spending were to grow steadily instead of following the fluctuations in petroleum revenues, this would help stabilize the economy and would also allow the state to accumulate a much larger fund to support the long-run demands for public services. Since some of the accumulated surplus would be used in later years to sustain long-run growth, this fund might more appropriately be termed a state growth fund rather than a permanent fund.

Two simulations (Cases 3 and 4) were carried out using a policy of steady growth in real state expenditures per capita; alternative growth rates of 6 and 8 percent a year were used. With either growth rate, the Alaska economy seems to end up in a stronger position, whether viewed in the aggregate or on a per capita basis. The overall growth rate is more stable, being lower in the earlier years and faster later on, and the employment rate is consistently higher. On a per capita basis, real disposable personal income in 1990 is $130–$250 higher (in 1967 prices) and real state spending is as much as $570 higher per person, depending on the growth rate used.

Another important effect of the steady expenditure growth policy is a much stronger state fiscal position. The balances accumulated in the state growth fund would

reach $10 billion with an 8 percent spending growth rate and more than $20 billion with 6 percent growth. In the previous case, it should be recalled these balances reached only $5 billion. By the late 1980's, it is likely that petroleum developments will no longer be driving the Alaska economy at such a rapid pace. At that time, the very large balances accumulated during the high growth years will leave the state government in a position to guide the economy onto a slower but more sustainable growth path.

It has been frequently suggested that some of the petroleum revenues be used to reduce state taxes. To examine the impact of this policy option, another policy simulation was run in which personal taxes were cut by 50 percent (Case 5). The direct cost of such a tax cut is $50 million in 1978, rising to $150 million by 1990. However, the tax cut also has indirect effects which accumulate over time and within a few years become significantly larger than the direct effects. The increase in economic activity induced by the tax cuts creates additional jobs, and this, together with the relative increase in disposable personal income, attracts additional migrants to Alaska. State spending then has to be increased to meet the needs of the larger population (it is assumed that the tax cut is not accompanied by a reduction in the average level of public services being provided).

With lower taxes and higher spending, the state is not able to accumulate trust fund balances nearly as rapidly as before. In fact, the growth fund at its maximum is less than half as large as the fund achieved in the absence of a tax cut. The interest earnings on the funds are cut by more than $200 million a year. The total cost to the state treasury (a total comprised of the direct tax cut, the loss of interest income, and the increase in spending) reaches $600 million a year by the end of the projection period. The cost is thus three to four times as large as the tax cut alone. Furthermore, because the potential gains tend to be dissipated by the increases in population, the increase in real disposable personal income per capita in 1990 is only $30, a gain of seven-tenths of 1 percent. Clearly, reducing income taxes is not an effective way of increasing the economic well-being of the average Alaskan.

III. Conclusions

There are several important policy conclusions derived from the Alaska model which seem to be generally applicable to a region involved in a resource development process. These conclusions are robust with respect to changes in the underlying assumptions concerning the magnitude of the development projects, and thus, seem to reflect the general nature of the process rather than the specifics of the Alaska situation:

Surges in state and local spending during development boom periods accentuate fluctuations in the regional economy and eventually lead to fiscal difficulties. When the pace of development subsides, the governments will have to cut back on public services, raise taxes, or draw heavily on their accumulated surpluses (if any). The problems will be particularly severe if the resource is exhaustible since, in that case, the resource activity will not merely level off but will decline.

An effective policy strategy to achieve a sustainable growth path would be to have state and local government expenditures grow at a steady rate based on long-run projections of revenues. When revenues grow at above average rates, the state government should accumulate surpluses that will be used in later years to support long-run growth.

Reducing personal income taxes is not an effective way of using revenues from resource developments to increase individual economic well-being. Induced migration quickly dissipates the gains in disposable personal income per capita. The increase in population also requires additional state spending and the total drain on the state treasury is several times as large as the tax cut alone.

REFERENCES

N. J. Glickman, "Son of 'The Specification of Regional Econometric Models,'" *Papers Reg. Sci. Assn.*, Nov. 1974, *32*, 155-77.

O. S. Goldsmith, "A Fiscal Model for Alaska: Structure and Policy Applications," paper presented at Annual Meeting, Western Economic Association, Anaheim, June 1977.

D. T. Kresge, (1974a) "Alaska Economic Growth 1961-1972," *Alaska Rev. Bus. Econ. Cond.*, Aug. 1974, *11*, 1-16.

———, (1974b) "Estimated Gross State Product for Alaska," *Alaska Rev. Bus. Econ. Cond.*, Apr. 1974, *11*, 1-17.

——— **et al.** *Issues in Alaska Development*, Seattle 1978.

D. A. Seiver, "Alaskan Economic Growth: A Regional Model with Induced Migration," paper presented at Annual Meeting, Regional Science Association, Cambridge, Nov. 1975.

Federal Energy Administration, *National Energy Outlook*, Washington 1976.

The Southwest: A Region under Stress

By LEE BROWN AND ALLEN V. KNEESE*

The states of New Mexico, Colorado, Utah, and Arizona form a region defined more by a wide range of common problems than by any shared institutional apparatus for solving those problems. The Southwest, as we term this region, is an arid area with its major watershed having an average annual precipitation of only sixteen inches. It is an area of high aesthetic quality with spectacular mountains, fertile valleys, and sweeping plains. Many of these vistas have been preserved in the national parks, monuments, forests, wilderness, and other federal and state reservations that have been established in the region.[1] It is a region whose history was formed largely by the force of three distinct cultures—Indian, Spanish, and Anglo—whose modern society retains this tricultural division. Although some blending of the separate cultures has occurred, there still exist numerous towns, villages, and settlements which are exclusively within one cultural tradition. It is a region of low population densities with most of the population concentrated in nodes of economic activity some distance from one another.

Of most immediate importance are the twin facts that the region contains many of the nation's poorest subsocieties, and the region is rich in natural resources, particularly fuels. An aggregate indicator of the first fact is found in the 1976 per capita income figures for Utah and New Mexico which stood respectively at 84 percent and 83 percent of the national level. More dramatic, however, is the economic condition of many of the Indian tribes within the region. As an example, the 1970 median family income of the 125,000 member Navajo tribe of New Mexico, Arizona, and Utah is but 32 percent of the national level. Further evidence of the disproportionate nature of income distribution is found in a Gini coefficient of .435 for Planning District #1[2] in New Mexico which is the northwest corner of the state and is the scene of a substantial portion of the energy development in the region. This planning district is the locus of a 2,155 megawatt power plant, a 690 megawatt power plant, numerous coal and uranium mines, and construction or proposals for construction amounting to approximately 3,500 more megawatts of electrical generating capacity and seven 250 million cubic feet per day coal gasification plants.

As for the natural resource reserves of the region, the four states contain an estimated 58 percent of the *U.S.* ten dollar uranium reserves,[3] over 23 billion short tons of coal in the reserve base estimates alone,[4] much of which is of low sulfur content, and the bulk of the nation's oil shale deposits. In addition to these energy fuels, the region is the location of substantial deposits of copper, molybdenum, iron, gold, silver, oil, natural gas, and other resources.

It is a striking feature of the region that to a large degree the subareas within the Southwest that contain the poorest populations are the same geographical areas that contain the bulk of the energy resources. This fact has, of course, given rise to increasing hopes and expectations that the development of these resources will provide the vehicle for substantial improvement in the economic welfare of the contiguous populations.

*Associate professor of economics and director of the Bureau of Business and Economic Research, University of New Mexico; and professor of economics, University of New Mexico, respectively.

[1]The federal government owns approximately 44 percent of Arizona, 36 percent of Colorado, 33 percent of New Mexico, and 66 percent of Utah.

[2]Furnished by Craig Bigler, consultant to the Four Corners Regional Commission.

[3]Taken from *1975 Uranium Statistics*. The market value of uranium oxide has, of course, far surpassed the ten dollar figure, but the regional share of the ten dollar reserves is representative.

[4]For more detail see the authors and M. D. Williams.

I. The Regional Issues

There are three regional issues which taken together encompass the bulk of the regional problems associated with economic development in the Southwest. They are 1) the environmental preservation issue, 2) the resource revenues issue, and 3) the water resource issue. Let us briefly describe the first two and then concentrate on the last problem.

A. *The Environmental Preservation Issue*

The Southwest has long been a principal battleground in the conflict between development and preservation. There have been battles over the once-proposed hydroelectric dams to be built in the Grand Canyon. More recently, the contest has shifted to the large coal-fired generating stations that have been proposed throughout the region. The 3,000 megawatt Kaiparowits facility in southern Utah was withdrawn in 1976 after a long, bitter, and costly regional and national debate. Clearly, this issue is not peculiar to the Southwest. Yet in this region the issue achieves a sharp focusing of alternatives which few other settings can match. There is no doubt that the high visibility and aesthetic quality of the region has long served the nation as a prime recreational asset. There is also no doubt that many of the societies which surround these resources are economically poor and that at the moment their only significant opportunity for economic improvement is related to resource development.

The Southwest has been no more successful than the rest of the nation in finding and implementing a policy solution to this issue, though the Navajos have recently adopted an innovative emissions charge system[5] which has long been advocated as an alternative to the existing regulatory scheme. In the meantime, each proposed development is individually contested at considerable cost to all parties. Until a developmental direction is firmly charted vis-à-vis the degree of environmental preservation and method for achieving it, this unresolved issue remains an obstacle to economic improvement.

B. *The Resource Revenues Issue*

As resource development in the region has increased, so has public interest in taxing that development. This interest, however, is not uniformly distributed throughout the region. New Mexico extracts the greatest revenues from resource extraction, realizing approximately $80 million in direct taxes in 1975. Colorado on the other hand did not even enact a severance tax until the 1977 legislative session. The different tax schedules reflect different taxing philosophies among the states and the different condition of the various state economies. Utah, for example, seems more concerned with insuring actual development of the resources and accordingly is wary of imposing taxes on resource extraction which might reduce the level of extraction. Though a similar concern is also expressed in New Mexico, it has in many instances been offset by a stronger concern that the state is not receiving an adequate share of the benefits from resource development. In our judgment, these differences in economic philosophy preclude any strongly integrated, cartel-like approach to resource taxation within the region, as has frequently been suggested. Nevertheless, the desire for immediate and sustained economic improvement is a strong force within the region, particularly in those resource-rich subareas which have lagged the remainder of the region in economic improvement. There is increasing recognition of the substantial current dependence of these subeconomies upon what are inevitably exhaustible resources. Correspondingly, there is a developing concern that actions be taken now to provide for tomorrow's economy. New Mexico voters recently approved a constitutional amendment designed to preserve a portion of resource tax revenues for public investment purposes rather than for current expenditures. The concern is erratically expressed, however, and fragile. The 1977 New

[5] *See Kneese and Charles L. Schultze.*

Mexico legislative session also rebated $100 million to New Mexico taxpayers rather than invest it developmentally within the state.

These factors will continue to exert strong upward pressure on resource taxation schedules within the region, with the limiting constraints provided by market forces, differing public attitudes towards taxation, and ultimately constitutional limitations on state taxing authority. The final levels of taxation and their effect upon developmental patterns within the region remain uncertain.

Finally, under this issue heading it should be pointed out that although the states of the Southwest have a substantial opportunity to accumulate resource tax dollars which can be used to promote long-term economic improvement, very little basic planning in the region has been directed towards meeting this need. The states have relied instead on conventional techniques such as tax advantages, foreign trade zones, and tourist promotions to enhance their economies. There has been little effort directed towards determining the region's long-term comparative advantages and designing an investment program aimed at cultivating those advantages. The region is currently accelerating economically under the stimulus of national and international energy forces and largely independent of efforts at the state level. It may be that the resulting economic improvement will be sufficient to dampen the forces seeking long-term planning before that long-term problem is solved. Moreover, there is evidence which suggests that the general improvement may largely bypass many of the subsocieties whose economic health is most in need of improvement.

II. The Water Issues

Of all western problems, water issues have received the most public attention in recent years. What formerly was perceived almost exclusively as a regional problem, except for frequent trips to the national treasury, has now acquired national status. A variety of factors have precipitated the change. Most noticeable are President Carter's recent budgetary restrictions on water development projects and the widespread drouth that has occupied much of the West. Less prominent, but no less significant, are the concerns over the adequacy of regional water supplies in meeting expanded energy demands and the increasing Indian assertions of rights to water in most of the water basins of the region.

In these circumstances it is important nationally as well as regionally to seek answers to the following questions. To what extent are water problems in the Southwest (as well as the West generally) a barrier to development within the region? To the extent that barriers do exist, what are the prospects for their removal? These are not simple questions, and the most we can do here is provide some perspective.

First, a brief background discussion is needed. All four states of the Southwest practice the legal doctrine of "prior appropriation" in allocating water supplies in contrast to the riparian doctrine common to the eastern United States. Succinctly stated, prior appropriation assigns a superior legal right to the earlier water user who puts the water to beneficial use.[6] As the regional population and economy have grown, increasing portions of both the region's surface and groundwater supplies have been appropriated. The Rio Grande in Colorado and New Mexico has been fully appropriated for a number of years. More importantly, the Colorado River, which is the largest watershed in the region in addition to being the locus of most of the region's energy reserves, is nearing full appropriation. Applications to put water to use already exceed reliable surface flow in the Colorado Basin in each of the states of the region.

In this context of full appropriation, let us enumerate and briefly examine four unresolved water issues within the region. They are the equity issue, the efficiency issue, the environmental quality issue, and the water development issue. Each of these terms denotes a generic class containing

[6]Riparian law in contrast gives the superior right to the user who is physically closer to the water body.

many subproblems. Principal examples of each will be listed for illustration.

A. *The Equity Issue*

Although the water basins have reached or are nearing full appropriation, the ownership of water rights remains highly unsettled. Though the federal government has tolerated, and on occasion endorsed, the prior appropriation doctrine practiced by the states, it has also in recent years promoted two other categories of rights—Indian and federal reserved[7]—which in many subareas are in strong conflict with rights established under the prior appropriation system. This unsettled question places a large cloud over the ownership and management of water in the region. Viewed abstractly the issue is but another instance of the international question of the distribution of benefits associated with resource development. In ordinary life within the region, it is another source of constant cultural and political friction.

B. *The Efficiency Issue*

As the region comes increasingly to accept the limited nature of the water supplies with which it must live, heavier reliance begins to be placed on the institutional procedures for transferring water rights from one owner and use to another. In many instances, these transfer procedures are cumbersome or outright inflexible. Arizona laws, for example, formed during an agriculturally dominant era, prohibit the separation of the water rights from the land to which they are appurtenant. Both must be bought or sold. The other three states are more flexible. Moreover, the large number of parties who may assert jurisdictional authority or other interest in opposition to a transfer further complicates the procedure. The results may be transaction costs for the engineering, legal, and administrative requirements to complete the transfer which amount to five or six times the value of the water right being transferred. Given the fundamental importance of water in an arid region, many of whose people are economically poor, the natural question that arises is the extent to which the water management system in the region should be refashioned to improve efficiency. There is much evidence indicating a growing strain on parts of the current management schemes, most noticeably in urban and energy resource areas.

C. *The Environmental Quality Issue*

The environmental questions illustrated above with air quality examples have their counterparts in the water arena also. These include conflicts over the increasing salinity of rivers[8] and in-stream uses. The latter problem concerns the extension of the definition of legally acceptable uses of water to include maintenance of streamflow as an aquatic habitat and for aesthetic reasons generally, an extension which is viewed by many established water users as a further erosion of their rights.

D. *The Water Development Issue*

It is this issue which was dramatically focused in early 1977 by President Carter's budgetary actions. Fundamentally, the issue involves several threads including preservation of scenic areas from dam building activities, the proportion of cost sharing between local, regional, and federal revenue sources, and the technical procedures of benefit-cost measurement. Water development projects, even if confined exclusively in the future to intraregional purposes, will receive close scrutiny. The era of mammoth western water developments has, not surprisingly, ended.

III. A Cautiously Optimistic Perspective on Water

It is, of course, impossible to detail here solutions to all of the individual problems

[7]Federal reserved rights describe rights to water in the region asserted by the federal government as necessary to fulfill the purposes of the many federal activities in the region.

[8]High salinity levels are particularly damaging to irrigated agriculture and have been the source of conflict between the United States and Mexico in the Colorado River Basin.

that fall into the four issue categories. In some cases, well-founded solutions have been intellectually worked out but have not found the necessary political adherents to be implemented. In other cases no workable solution has yet been found. Let us, however, briefly paint a general perspective on water as a problem for development in the Southwest.

We find the region's water institutions to be its most sophisticated creations. What is more important, these institutions are proving to be adaptive to changing conditions although the process is slow and almost always painful to certain segments of society. New electrical generating stations are introducing hybrid cooling technologies with reductions in water consumption up to 75 percent. More importantly, markets for water rights have emerged in three of the states, which, although rudimentary in many ways, are providing price signals reflecting the relative scarcity of water in different basins. Still, the problems are difficult and the conflicts harsh. The drouth of the past several years has provided a severe test of the region's water institutions, and the strain is evident. In an abstract summary it is almost impossible to convey the strength of the emotions that are attached to water in the West. Whole novels have been written with water conflict as their central theme (see, for example, John Nichols). Water rights have become a symbol to many Indians of their aspirations to political sovereignty and improved economic condition. Practitioners of the ancient technology of irrigated agriculture express dismay at the prospect of water right transfers from agriculture to "more valuable" economic uses. Environmentalists ardently defend against any further development of water projects which are injurious to the natural environment.

In this emotion laden context there will be no easy or painless resolutions of the water issues. However, with a strong measure of abstraction it becomes possible to cautiously assert a measure of optimism. There are three reasons for this optimistically shaded perspective. First, new water users in the region are proving to be strongly adaptable to water conserving technologies as indicated above for the energy case. Second, there is the healthy evolution of the region's institutions as indicated by the development of the rudimentary market for water rights. Although it must be expected that such a market will always be strongly tempered by public control, its existence does increase the flexibility and efficiency of water use in the region. Finally, the third reason lies with the increasing sophistication and forcefulness with which all parties to the issues assert their interests. Although in the short term this assertiveness may be expected to increase the number of points of conflict and the clamor that surrounds them, in the long run it is healthy. No longer will important interests be passed over only to be reasserted at a later date when the issues had seemingly been settled.

Each of these three factors requires large amounts of time to become effective. It is this aspect which injects a cautionary note into the perspective for the pace of development within the region is not within the region's control. At the pace of energy growth originally forecast for the region following the events of 1973 that time would not have been available. Only makeshift solutions could have been applied. As the pace has slowed, the more sanguine perspective has become feasible. Given time, the region may yet find a way to reduce or even eliminate water as a barrier to economic improvement.

REFERENCES

L. Brown, A. Kneese, and M. D. Williams, "The Southwest: A Region Under Stress," unpublished paper, Univ. New Mexico 1976.

Allen V. Kneese and Charles L. Schultze, *Pollution, Prices and Public Policy*, Washington 1975.

John Nichols, *The Milagro Beanfield War*, New York 1974.

Kerr-McGee Nuclear Corporation, *1975 Uranium Statistics*, Oklahoma City 1975.

The New England States and their Economic Future: Some Implications of a Changing Industrial Environment

By JOHN R. MEYER AND ROBERT A. LEONE*

Perhaps the most striking feature of the New England economy is that it is different—not only from the rest of the nation, but from the rest of the northeastern United States as well. New England's main departure from the national norm is, of course, relatively slow growth; it is conventional to describe New England as "mature"—economically, industrially, and maybe even demographically.

This maturity manifests itself in many ways. While aggregate personal income in the United States expanded at an average annual rate of 4.1 percent between 1960 and 1975, New England expanded at a rate of 3.6 percent per annum. Further, total manufacturing employment in Massachusetts and Rhode Island is only slightly higher today than in 1914. In the recent recovery from recession, New England lagged well behind the rest of the United States in expansion of total employment but nevertheless recorded some of the sharpest declines in unemployment rates so that New England unemployment is now near the national average even though it was much higher at the depth of the 1975 recession. The secret, of course, to New England's relatively rapid unemployment decline is slow workforce growth, as expected in a mature economy.

New England's differences from the rest of the Northeast are perhaps less obvious and certainly less well known. It is fashionable today to speak in very broad terms of "frostbelt" vs. "sunbelt" and to suggest that public policy should modulate differences in growth among the different sections of the country. The reality, though, is that aggregate figures for large regions of the country hide a good deal of internal diversity. Thus, New England not only seems to be doing better than conventional frostbelt wisdom would suggest, but its immediate prospects also appear more favorable than those of the mid-Atlantic states and probably much of the Midwest as well (see Benjamin Stevens and Glinnis Trainer). Even in the recent past, as between 1960 and 1976, when New England's aggregate personal income was growing 3.6 percent per year, the states of New York, New Jersey, Pennsylvania, Maryland, and Delaware had a combined average annual compound growth rate of only 3.3 percent. Similarly, a "shift-share" analysis has indicated that the entire Northeast (by virtue of a favorable industry mix) should have been in a position to gain in share of *U.S.* jobs throughout the 1960's. The New England states (except for extreme northern Maine) have indeed done as expected. Large areas of the remaining Northeast, however, have experienced significant competitive shifts or losses (see Richard Olsen).

To a considerable extent, in fact, any New England "success" in recent years may have been at the expense of its immediate neighbors. New England production costs perhaps have not been as low as in much of the Southeast in recent years, but they apparently have been competitive with the Middle Atlantic, and especially New York City. In fact, total manufacturing costs in several industries (for example, ordnance, primary metals, fabricated metals, nonelectrical machinery, transportation equipment, paper and printing) have been lower recently in Massachusetts (probably the highest cost New England

*Harvard University. The Economic Development Administration of the U.S. Department of Commerce and the 1907 Foundation provided financial support for this research.

state) than in the United States as a whole (see George Treyz). Indeed, New England labor costs, when adjusted for skill and industry composition, may be lower than in the sunbelt.[1] Northeastern market access, moreover, would help offset any residual wage disadvantage that might still exist.

Whatever the explanation, New England has acquired new manufacturing employment in recent years because of plants moving from the greater New York City area to New England. Specifically, of thirty-nine manufacturing establishments which can be identified as moving into New England between 1967 and 1971, twenty-six came from the New York Standard Consolidated Area (*SCA*);[2] by contrast, of twenty-seven establishments identified as moving out of New England during that period, only seven moved into the New York *SCA*.[3] For twenty other plants identifiable as emigrants from New England in these years very little pattern is discernible since they had destinations in thirteen different states. Measured in jobs (1971 employment), 1,139 left New England[4] while 3,507[5] moved in. Of these immigrants, 2,517 jobs had their origin in the New York *SCA*. Thus, except for transfers from New York, New England appears to have neither gained nor lost manufacturing employment due to establishment relocations in these years. As one Yankee wit has observed: "New England is to the New York City area as New Hampshire is to Massachusetts."

An obvious question is why New England's recent performance and prospects seem so favorable relative to the rest of the frostbelt.[6] The most obvious, and probably fundamental, explanation of New England's comparative state of economic bliss is simply that it has already suffered for so long. In 1914, shoes, leather, and textiles represented 50 percent or so of New England manufacturing jobs; today they represent only about 10 percent. The transition in New England's industrial structure has been long and difficult for workers, cities and towns, as well as investors (see Roger Schmenner) but as a consequence, the New England economy is perhaps closer to an equilibrium in its factor cost and structural relationships than other regional economies of the United States. For example, the mid-Atlantic states have only recently been subjected to sharply negative external changes, like those that befell New England earlier.

This equilibrating process, as described long ago by George Borts and Jerome Stein, has capital and labor flowing in response to differential returns in an open economy with well-defined production and demand characteristics. In reality, technology and markets are continually changing so that the relative economic attractiveness of different geographic areas is also evolving. Whether these forces will offset or accentuate the "natural" adjustment process is very much an open question. For New England, though, these

[1]Frank Morris, President of the Federal Reserve Bank of Boston, indicated during a talk at the Joint Center for Urban Studies of Harvard and M.I.T., that this would be the conclusion of a forthcoming study by the bank.

[2]These figures were developed by the authors from comparison of Dun & Bradstreet Market Indicator tapes for these years and *exclude* thirty-seven establishments which moved from the New York metropolitan area to Fairfield County, Connecticut.

[3]This *excludes* four establishments which moved from Fairfield County to the New York area.

[4]This figure is probably very low. Three emigrating plants which had total employment of 525 workers in 1967 failed to report employment in 1971. For those reporting employment in both years, the 1967 employment total was 4,178 versus 1,117 in 1971; thus, plants which left New England appeared to have contracted in size after the move.

[5]In those plants reporting employment in both 1967 and 1971, the 1967 employment total was 1,604 versus a 1971 total of 3,454; thus plants which moved into New England appear to have expanded in size after the move.

[6]The hypotheses outlined here emerge from several sources: for example, work done by colleagues at Harvard, M.I.T., and other New England universities and colleges; studies completed at the Boston Federal Reserve Bank; and some of our own preliminary investigations of "industrial demographics" and related topics (see Benjamin Chinitz and Treyz et al.). In particular, many of these thoughts emerged at a Conference on the Future of the New England Economy held in January 1977 at the Joint Center for Urban Studies of M.I.T. and Harvard.

external or exogenous changes in the general business environment have been of late largely favorable to the adjustment processes otherwise at work. Specifically, they have 1) attenuated many historical disadvantages of New England locations; 2) accentuated certain long-standing advantages; or 3) created possibilities for New England industry that were simply not there before.

The interplay between economic maturity and the development, or disappearance, of regional comparative advantage is perhaps best illustrated by infrastructure investments. Due to its maturity, New England has in place a physical infrastructure—ports, roads, hospitals, schools, etc.—which when new was a major source of the region's economic strength. Over time, however, as technology evolved and relative factor prices changed, this old infrastructure became as much a liability as an asset. Above all else, it was often cheaper to develop new infrastructure from "scratch" elsewhere. As a consequence, New England paid a price, so to speak, for being early in the development process: newer cities and regions unencumbered with old assets could meet their infrastructure requirements at a lower cost than New England.

Today, for a variety of reasons such as changing factor prices, general inflation, government regulations, and in some instances, diminishing returns, new infrastructure appears to be of an increasing cost character as often as not (see Kent Anderson and James de Haven). In particular, technological gains are no longer quite as certain to offset factor-cost inflation. Accordingly, quasi rents can accrue to facilities already in place. The proportional extent of such windfalls is greatest, of course, for those areas with the largest stock of low cost capacity already in place relative to demands for the services of such facilities. Additionally, it is often cheaper today for environmental and other reasons to "round out" or expand capacity at existing sites with established facilities than it is to undertake entirely new investments. When slow growth is coupled with a relatively low cost of expanding existing infrastructure, any remaining infrastructure disadvantage of mature regions can be substantially attenuated.

In general, the marginal costs of providing additional basic support activities in different regions appear increasingly alike. The phenomenon is perhaps most aptly illustrated by electricity. In the post-World War II period (say, until about 1965), the real cost of electricity in the United States was falling (see Irvin Bupp). New increments to capacity needed to satisfy the almost constant 7 percent or so compound annual growth rate in consumption came on line at a cost cheaper than that of the existing capacity. In such an environment, the average historical cost-pricing methods of the utility regulations were to everyone's advantage: rates to consumers fell over time while on the margin there was always a healthy incentive to expand supply.

In the late 1960's and early 1970's a variety of factors changed this situation. For one, the capital cost of new capacity grew substantially due to expensive antipollution requirements and a lengthening of the time between the decision to build and actual plant start-ups. Furthermore, most low cost sources of power and the simpler productivity improvements had been exhausted. With the incremental costs of new capacity greatly exceeding the average historical cost of existing plants, the "old" regulatory rules based on average cost pricing impart a severe bias against expansion of capacity. In this environment, however, average costs and rates for New England utilities should rise less rapidly than in other regions, since New England utilities have little low cost power in their existing base. The situation of the New England utilities is further helped by their effective rate of capacity utilization in 1975 being 4 percentage points below the national average (see Edison Electric Institute); this "excess capacity" was apparently created by a surge in nuclear installations in New England in the last few years since as recently as 1965

New England utilities experienced a slightly higher percentage of capacity utilization than the national average.

Of course, regulators might adopt marginal cost pricing in response to the pressures created by rising unit costs. If so, the increase in rates will be far less traumatic for New England consumers than for those located elsewhere. For example, if electricity were to be charged at incremental costs in both New England and the Pacific Northwest, prices might double in New England but increase fourfold in the Pacific Northwest. And since New England users long ago confronted high costs, the region's capital stock has evolved accordingly, further mitigating the trauma generated by any increase in energy costs. A most striking statistic about New England industry is how relatively little energy it seems to consume; in 1967 industrial electricity consumed per $1,000 of value-added was only 957.2 kilowatt hours in New England versus 2,186.3 on average for the United States as a whole.

Higher energy prices may also attenuate the disadvantage inherent in New England's deficiency of natural resources. This will be particularly true if relative commodity prices gravitate upward over time. No one would predict that New England will soon become a major reservoir for natural resources, but the mere existence of exploratory efforts to locate coal and oil is at least suggestive. At any rate, if the quality of conventional raw material sources deteriorates, the attractiveness of recycling should increase. A rise in energy prices would reinforce this trend since recycled raw materials normally require less energy per unit of output. Given New England's high density and affluency, the region is an obvious source for paper, aluminum, and steel, among other recyclable materials.

In general, recent and prospective factor price changes have the potential of creating certain industrial opportunities in locations where they did not exist before. Evidence in the case of steel for such a shift is in fact fairly impressive. Recent experience suggests that small-scale minimills can produce steel for about $175/ton, or about $50–$100/ton lower than most other domestic producers (see David Santry). These savings are in part due to the fact that new minimill capacity incurs capital costs of about $100/ton—a sum which is about one-tenth that of new large-scale integrated mills. The economics of steel production thus indicate opportunities for new small-scale, market-oriented manufacturers, even in an industry increasingly impacted by foreign (and some say, heavily subsidized) imports. By contrast, many large-scale energy-intensive steel manufacturing facilities to be found outside New England may suffer even more contraction as they adjust to the changing industrial environment. To illustrate, rising energy prices are not likely to cause the steel industry, as conventionally established in large-scale open hearth and basic oxygen process mills, to leave Gary, Indiana for Worcester, Massachusetts, but they may encourage the relative growth of less energy-intensive, direct reduction minimills which are attracted to high-density sources of steel scrap, such as Worcester.

Other recent changes in the industrial environment may also create some unexpected advantages for New England. Consider as a specific case the impact of federal water pollution controls. In general, water pollution controls mainly impact industries that are also heavy users of energy (for example, textiles, petroleum, chemicals, steel, pulp and paper, metal plating); activities which for the most part were never well suited to New England or have long since fled. One major exception would be tissue paper manufacturing, an industry of considerable importance to New England today. On the basis of cost estimates for sixty-four tissue mills in the United States, eleven of thirteen New England mills should experience a deterioration in their *relative* competitive position (measured by total returns to invested capital) as a result of the Environmental Protection Agency's (*EPA*) 1977 pollution abatement standards (see John Jackson

and Leone). The other two mills should recover their added costs and earn a reasonable return on their investment in pollution control devices. On average, *EPA*'s 1977 standards thus create a sizeable economic disadvantage for New England producers. However, as a side effect of complying with 1977 standards, New England producers face relatively *low* incremental costs to achieve the tighter standards scheduled for 1983; twelve of the thirteen mills would benefit and the position of the thirteenth would remain unchanged.[7] Meanwhile, all existing tissue manufacturers will actually have their economic lives extended since they face lower variable costs (including the *full* costs of pollution controls) than new plants, whose costs are raised significantly by *EPA* compliance requirements. This, in effect, buys time for New England mills, which might otherwise have been retired in the near future.

In sum, the environment in which industries operate is constantly changing. These changes create relative economic advantages and disadvantages within industries and across regions over time that can either reinforce or obstruct the natural tendency in an open market economy for regional differences to be attenuated. In the particular case of New England, these external changes recently have reinforced the equilibrating processes, thus facilitating some revival of the New England economy. This combination may not spell boom for the New England economy, but it certainly does not foretell disaster.

There are perhaps lessons in the New England experience for others as well. If, indeed, the current "baby dearth" proves durable, some other regions of the United States will make the transition from demographic growth to stability sometime in the not too distant future. Whether this is accomplished by economic stagnation depends of course on a host of other factors, not the least of which is the general state of the *U.S.* economy. The evidence suggests that the need for transition in industrial structure is now seeping southwestward from New England. Indeed the steel industry will be to the 1970's what textiles were to the 1950's. Whether these other regions will adjust with the same, lesser, or greater facility than New England remains to be determined since, as the central argument of this paper points out, much depends on the specific forces at work conditioning both relative and absolute costs, not only within the regions, but without as well. Perhaps the only certainty is that as industrial maturity spreads outward from its original locus in New England, public policy will be much more sensitive and attuned to the problems of regional economic decline. Indeed, one of the ironies of current public policy as viewed from a New England perspective is that allocation formulas for public funds and other federal aids for regional economic transitions are becoming more favorable and available to New England just as the worst of its own transition problems may be at an end. It remains to be seen, of course, whether federal assistance for troubled regional economies will accelerate or hinder these transitions since public policy, like any other exogenous influence, can either obstruct or reinforce the process.

[7]It is therefore somewhat ironical that the portions of the industry hardest hit by the 1977 standards have been attempting to defer application of the 1983 standards to 1984, while also making them less stringent. In mid-November 1977, a House-Senate conference committee agreed to less stringent 1983 standards, but this agreement has yet to be reflected in law.

REFERENCES

Kent P. Anderson and James C. de Haven, *Long Run Marginal Costs of Energy*, Santa Monica 1975.

George H. Borts and Jerome L. Stein, *Economic Growth in a Free Market*, New York 1964.

I. C. Bupp, "The Electric Utility Industry: Reference Note," Intercollegiate Case Clearing House, Boston 1975.

B. Chinitz, *The Decline of New York in the 1970's: A Demographic, Economic and Fiscal Analysis*, Binghamton 1977, 57–78.

J. Jackson and R. A. Leone, "The Political Economy of Federal Regulatory Ac-

tivity," mimeo., Harvard Bus. Sch., Boston 1977.

Richard J. Olsen, *Multiregion: A Socioeconomic Computer Model for Labor Market Forecasting*, Schloss Laxenburg 1976.

D. G. Santry, "Nucor: One Winner in Troubled Steel," *Bus. Week*, Nov. 21, 1977, 104.

R. Schmenner, "The Manufacturing Location Decision," mimeo., Harvard Bus. Sch., Boston 1977.

B. H. Stevens and G. A. Trainer, "Further Thoughts on the Prospects for Growth In Massachusetts Manufacturing," in *Massachusetts Business and Economic Report*, *4*, No. 4, Amherst 1977.

G. Treyz, "The Massachusetts Economic Policy Analysis Model," mimeo., Univ. Mass. 1977.

———, A. F. Friedlaender, and B. H. Stevens, "A Regional Economic Policy Simulation Model," mimeo., Cambridge, Mass. June 1977

Edison Electric Institute, *EEI Statistical Year Book of the Electric Utility Industry for 1975*, New York 1976.

DISCUSSION

BENJAMIN CHINITZ, State University of New York-Binghamton: The paper on the Southwest deals briefly with the issue of resource revenues, the main issue in the Alaska paper. Is the latter model relevant to the former? While it probably would be much harder in the case of the Southwest to project the flow of revenues and the impact of fiscal policy on migration and local economic activity, policymakers in the Southwest would do well to study the Alaska model and its findings.

The paper on New England and the paper on the Southwest dramatize the superficial nature of the popular sunbelt/snowbelt dichotomy. John Meyer and Robert Leone call attention to the significant variations within the snowbelt, and Lee Brown and Allan Kneese remind us that there is still abject poverty and economic anemia in sections of the southwest.

The very first sentence in the Southwest paper troubled me personally because it states flatly that the four states have no "shared institutional apparatus for solving those problems." I have spent the last four months evaluating regional commissions, one of which is the Four Corners Regional Commission embracing these four states and Nevada. I am shocked by the apparent and assumed impotence of the Commission in this context.

I have two problems with the Alaska model. I don't see why the later decline in public sector expenditures would be viewed as a problem if in the earlier period a substantial amount were devoted to public capital expenditures which would generate a stream of services into the future. Thus, the "real" flow of public services could be held constant, even if money outlays were diminished. Isn't this simply another form of savings which the authors want to encourage in the period of high revenues?

My second complaint is an old one (with me) and it may not survive a careful study of the model. I have in mind the prior once-and-for-all definition of exogenous and endogenous sectors. I have long argued that the regional multiplier is variable because the proportion of new demands which is satisfied locally grows over time in the process of growth. Even in Alaska I would expect some import substitution as the local market expands. This would provide a built-in stability which the authors seem to have ignored.

Meyer and Leone are on target in recognizing that part of New England's favorable performance is the other side of the coin, that is, it reflects an outflow from the New York region. My own research suggests that this phenomenon is even more pronounced in the service sector than in manufacturing. Even more importantly, if part of New England's strength is a function of these intranortheastern shifts, then we cannot be as confident about New England's future as the authors would like us to be because the outflow could be arrested by the same equilibrating forces which have been advantageous to New England.

WALTER ISARD, University of Pennsylvania: These papers pose a challenge to me and to all regional scientists concerned with the study of a multiregion system. The challenge centers around the development of an operational framework for deepening, in the study of an individual region, the analysis and projection of its connections with every other region in a system, for example, the *U.S.* system. Such an operational framework would permit a more systematic, and perhaps comprehensive, approach to examine changes, via the interregional linkage network, or more formally, via subsystem coupling functions. These changes could relate to improvement or deterioration of a region's interregional competitive position, or position of complementarity.

To be specific, all three papers refer to the energy sector and its critical role in regional development. Suppose then, we examine shifts in interregional patterns and thus interregional connections associated with the three major industrial consumers

of energy resources and related feedstocks. These are the iron and steel, the petrochemicals, and the aluminum electrolytic products sectors. I do not have space to present the full projections to year 2000 of the regional distribution for these three industries—projections based on comparative cost analysis and several iterations of population and market magnitudes using models that I and my associates have developed. However, a brief outline is sufficient.

In the case of iron and steel, we see relatively large increases in production in the Southwest, Pacific, and East North Central. There is very little absolute increase in New England and marked relative decline in the Middle Atlantic. On this count, and allowing for multiplier and subsequent agglomeration effects, we see a definite deterioration in New England's interregional competitive position and a definite pressure for use of the Southwest's coal resources. In the case of petrochemicals, we foresee a huge growth in the Gulf, with modest increases in the Midwest, South Atlantic, and West Coast. Again this represents a significant pull away from New England, directly and indirectly, and a definite increase in income, employment and market masses in the Gulf Coast and areas adjacent to the Southwest. In the case of aluminum, we see big increases in the South Atlantic, East South Central, and Pacific regions with the largest increase in the Southwest (based on coal generated power). We also foresee significant capacity in Alaska, double that currently in the Middle Atlantic and New England.

In sum, I see a continuing decrease in relative accessibility to national markets confronting New England (in terms of masses directly and indirectly associated with these three major industrial sectors). Now in their excellent and imaginative study of New England, John Meyer and Robert Leone have looked for key changes in the interregional connections of New England. They recognized, for example, steel as a key sector almost entirely outside New England whose health elsewhere is important for New England. However, they have not nor can they be expected at this stage to pursue in-depth analyses of regional distributions of key sectors and interregional connections—analyses which could provide extremely valuable materials to supplement their intensive study of New England and enable them to achieve an even more sound and thorough evaluation of the future of New England.

With respect to Lee Brown and Allan Kneese's paper of deep-probing research on the Southwest, it is clear, as they intimate, that some of the actors (institutions) involved in developing resource and other policy for the Southwest lie outside the Southwest. Thus, in suggesting and evaluating alternative policies, it becomes extremely important to anticipate the likely patterns of interregional connections, so as to identify more clearly what the interests of the actors outside the region might be. At the present stage they, too, cannot be expected to have conducted research on these likely patterns.

Lastly, with respect to David Kresge and Daniel Seiver's promising policy study for a region under *rapid* development, analysis of the range of likely patterns of interregional interconnections could provide invaluable materials for extending their already-penetrating policy analysis to a consideration of alternative structures (for example, one involving significant aluminum production) that might emerge for the Alaskan economy.

In sum, we have three excellent research studies. Each, however, seems in a position to profit greatly from intensive analysis of likely changes in interregional connections as the *U.S.* system of regions develop—analysis which could also lead to a set of more consistent regional policies. I urge that such analysis be initiated to complement individual region studies.

Energy Policy and U.S. Economic Growth

By EDWARD A. HUDSON AND DALE W. JORGENSON*

The purpose of this paper is to quantify the impact of alternative energy policies on future energy prices, energy utilization, and the structure and growth of the *U.S.* economy. The nature and magnitude of these interrelationships between energy and the economy have assumed great importance in view of the rise in world petroleum prices since 1973. Federal government price controls have prevented the full impact of these oil price rises from being felt by energy consumers but policy measures currently under consideration would not only raise domestic crude energy prices to world levels but would also raise delivered energy prices above world levels. These increases in energy prices will lead to a reduction in the growth of energy consumption. They can also have an important impact on future *U.S.* economic growth.

I. Econometric and Process Analysis Models

A satisfactory framework for analysis of the effect of alternative energy policies requires an approach that encompasses both process analysis and econometrics. Process analysis provides for a detailed characterization of technology for energy conversion and energy utilization and permits the analysis of effects of introducing new energy technologies. Econometrics provides for the incorporation of behavioral and technical responses of patterns of production and consumption to alternative energy prices and permits an analysis of the impact of energy prices on the demand for energy, nonenergy intermediate goods, capital services, and labor services. By representing energy sector transactions in physical terms we can provide a link to process analysis models. By representing these transactions in economic terms, in current and constant prices, we can provide a link to econometric models. By using both forms for representing energy transactions, process analysis and econometric modeling can be combined within the same framework.[1]

The first component of this framework is the Long Term Interindustry Transactions Model (*LITM*) developed by the authors.[2] In *LITM*, the technology of each producing sector is represented by an econometric model based on the price possibility frontier, giving the supply price of output as a function of the prices of primary and intermediate inputs and the level of productivity.[3] Technical coefficients giving primary and intermediate inputs per unit of output of the sector as functions of prices and productivity can be derived from the price possibility frontier. Given the level of output of the sector, the technical coefficients determine demand for intermediate and primary inputs into production. The preferences of the household sector are represented by an econometric model determining demand for consumption goods, demand for leisure, and supply of saving.[4]

*Data Resources, Inc. and Harvard University, respectively.

[1] Annual interindustry accounts for the United States for the period 1947–71 in physical and economic terms have been prepared by Jack Faucett Associates. Interindustry accounts for the year 1967 have been compiled for a more detailed industry breakdown by Clark Bullard and R. A. Herendeen (1973a,b).

[2] The model of interindustry transactions is described by the authors (1973, 1974, 1977).

[3] The model of producer behavior is described by Ernst Berndt and Jorgenson. See also Laurits Christensen, Jorgenson, and Lawrence Lau (1973) and Berndt and D. O. Wood (1975). A comparison of econometric and process analysis models of energy consumption is given by Berndt and Wood (1977).

[4] The model of consumer behavior is described by Jorgenson. See also Christensen, Jorgenson, and Lau (1975), and Jorgenson and Lau.

The second component of our modeling framework is the Time-Phased Energy System Optimization Model (*TESOM*) developed at the Brookhaven National Laboratory by K. C. Hoffman and associates.[5] This model is based on the Reference Energy System, which provides a physical representation of technologies, energy flows, and conversion efficiencies.[6] Within each period *TESOM* allocates energy supplies to energy demands so as to minimize cost. This is formulated as a linear programming model of the transportation type. Given levels of demand for energy services, available supplies of energy resources and conversion capacities, conversion efficiency and capital and operating costs of utilizing technologies, the energy sector optimization model determines the set of energy conversion activities and operating levels that minimizes total cost. Between periods, investment changes the capacities of the conversion processes.

The combined *LITM-TESOM* framework models interindustry transactions as a result of a dynamic general equilibrium of the *U.S.* economy.[7] In each period the relative prices of all commodities are determined by balance between demand and supply; technical coefficients for inputs of intermediate goods and primary factors of production are determined simultaneously; final demands are calculated and these, with the technical coefficients, determine demands for the output of each sector of the economy and for primary factors of production. In each period the supply of capital is fixed by past investments. Variations in demand for capital services affect the price but not the quantity of these services. Similarly, the available labor time in each period is fixed by past demographic developments. Variations in demand for labor time by the producing sectors and by the household sector for consumption in the form of leisure affect the price of labor and the allocation of labor time between market and nonmarket activity. Finally, the supply of saving by the household sector must be balanced by final demand for investment by the producing sectors. Dynamic adjustment to changes in energy policy is modeled by tracing through the impact of investment on future levels of capital stock.[8]

[5]The Brookhaven optimization models are described by Hoffman; Hoffman and E. A. Cherniavsky; Cherniavsky. A comparison of these models and alternative process analysis models of the *U.S.* energy sector is given by Tjalling Koopmans in L. Kantorovich and Koopmans.

[6]The Reference Energy System is described in M. Beller.

[7]The integration of the Hudson-Jorgenson model with the Brookhaven Energy System Optimization Model (*BESOM*) is discussed in detail by Hoffman and Jorgenson. A dynamic version of our econometric model of interindustry transactions is discussed in detail by the authors (1977). A dynamic version of *BESOM* has been developed by W. Marcuse et al. A comparison of the combined model with alternative models for analyzing the relationship of energy and economic growth is given by William Hogan (1977b).

II. Alternative Energy Policies

The starting point for our analysis of the impact of alternative energy policies is a base case projection of future energy and economic growth with no change in energy policy. We assume that any quantity of petroleum imports is available at the world price, where the world price of petroleum rises at a rate of 1 percent per year relative to the rate of growth of the *U.S. GNP* price deflator. The annual rate of growth or real *GNP* is projected to average 3.2 percent from 1977 to 2000. This growth rate is considerably below the 3.8 percent average annual growth experienced between 1950 and 1973. The decline is partly due to a reduction in population and labor force growth, and partly due to a reduction in productivity growth resulting from higher energy prices. Primary energy input is projected to rise from 76 quadrillion Btu in 1977 to 139 in 2000, an average annual growth rate of 2.6 percent, also below the average annual growth rate between 1950 and 1973 which was 3.5 percent. Part of the reduction is due to decreased economic growth and part is due to the conservation induced by regulation and by a continuing increase in real energy prices.

[8]A theoretical analysis of this dynamic adjustment process is presented by Hogan (1977a).

TABLE 1—ENERGY PRICES AND QUANTITIES IN THE YEAR 2000

	Base Case	Policy			
		1	2	3	4
Prices[a]					
Coal	1.64	1.71 (4.3)	1.71 (4.3)	4.07 (148)	6.12 (273)
Refined Petroleum	4.47	4.79 (7.2)	5.84 (30.6)	8.77 (96)	13.67 (206)
Natural Gas	3.47	4.06 (17.0)	5.12 (47.6)	8.33 (140)	12.99 (274)
Electricity	10.54	10.89 (3.3)	10.99 (4.3)	18.30 (74)	27.11 (157)
Average Price of Delivered Energy	5.21	5.30 (1.7)	5.86 (12.5)	9.78 (88)	14.97 (187)
Quantities[b]					
Coal	32.7	40.2 (23)	41.3 (26)	27.3 (−17)	20.0 (−39)
Petroleum	48.8	35.1 (−28)	28.2 (−42)	24.0 (−51)	19.4 (−60)
Natural Gas	19.0	17.4 (−8)	14.9 (−22)	15.0 (−21)	13.2 (−31)
Nuclear	28.5	23.7 (−17)	21.7 (−24)	18.1 (−36)	13.0 (−54)
Other	9.6	10.2 (6)	10.2 (6)	5.5 (−43)	4.3 (−55)
TOTAL	138.5	126.6 (−9)	116.3 (−16)	89.9 (−35)	69.9 (−50)
Imports as Percent of Total Input	20.4	10.7	3.0	1.9	0.0

[a]Delivered prices in $1975/million Btu; percentage difference from base case levels in parentheses.
[b]Primary energy input in quadrillion Btu; percentage difference from base case levels in parentheses.

We consider four sets of energy policies intended to reduce energy growth and to reduce dependence on imported energy sources:

POLICY 1: Taxes are imposed on *U.S.* petroleum production to bring domestic petroleum prices to world levels; natural gas prices are increased but price controls are retained; energy conservation is stimulated by taxes on use of oil and gas in industry, restriction of oil and gas use by electric utilities, subsidies for insulation of structures, and mandatory performance standards for energy-using applicances.

POLICY 2: The measures included in Policy 1 are combined with tariffs on imported oil rising to $7.00/barrel in 2000 and with corresponding taxes on natural gas.

POLICY 3: Policy 2 is combined with excise taxes on delivered energy sufficient to reduce total primary energy input in 2000 to 90 quadrillion Btu.

POLICY 4: Policy 2 is combined with excise taxes on delivered energy sufficient to reduce total primary input in 2000 to 70 quadrillion Btu.

Delivered energy prices under each policy in 2000 are presented in Table 1. Under Policy 1, the average price of delivered energy is only 1.7 percent higher than the base case, while in Policy 4 the increase is 187 percent. Extensive nonprice conservation measures explain the small indicated price rise between the base case and Policy 1. The average annual rate of increase in real energy prices between 1977 and 2000 is 1.6 percent in the base case and Policy 1, 2.1 percent in Policy 2, 4.4 percent in Policy 3, and 6.3 percent in Policy 4. The policies also change the structure of energy prices with petroleum and natural gas becoming more expensive relative to other fuels. The level and pattern of primary energy input in 2000 for each policy scenario are also presented in Table 1. Policies 1 and 2 have similar impacts on energy input, leading to reductions of 9 percent and 16 percent, respectively, from the base case; Policies 3 and 4 are more drastic, involving reductions of 35 percent and 50 percent. The policies change the pattern of energy input, decreasing the relative importance of petroleum and increasing that of coal. The reduced demand for petroleum and natural gas leads to large reductions in the degree of dependence on imported energy.

Increases in energy prices and energy conservation measures have a widespread and significant impact on the structure and

TABLE 2—SECTORAL PRICES AND QUANTITIES IN THE YEAR 2000

	Base Case	Policy 1	Policy 2	Policy 3	Policy 4
Prices[a]					
Agriculture		1.1	1.3	2.3	5.5
Manufacturing		.1	−.2	1.1	5.9
Transportation		3.1	4.6	7.8	13.3
Services		−.3	−.9	.6	2.2
Energy		1.7	12.5	87.7	187.3
Final Output		.7	1.0	5.1	11.0
Quantities[b]					
Agriculture	8.4	8.3	8.2	8.2	8.2
Manufacturing	30.5	30.5	30.4	30.3	29.6
Transportation	3.6	3.5	3.4	3.3	3.2
Services	54.1	54.7	55.1	56.0	57.2
Energy	3.4	3.0	2.8	2.2	1.7
Quantities[c]					
Agriculture	8.6	8.5	8.5	8.4	8.2
Manufacturing	35.9	35.9	35.8	36.0	36.0
Transportation	5.0	4.9	4.9	4.8	4.7
Services	45.7	46.1	46.5	47.4	47.8
Energy	4.8	4.6	4.3	3.4	3.2

[a]Percentage change in output price index from base case.
[b]Percentage composition of real final demand.
[c]Percentage composition of real output.

growth of the economy. The initial step in this process of adjustment is the restructuring of relative prices of goods and services. Table 2 shows the changes in output prices in 2000. Output prices rise in line with the energy content of each type of product—delivered energy prices rise the most, services prices the least. Producers respond to higher energy prices and energy conservation regulations by altering input patterns so as to minimize unit costs in the face of the new price structure, subject to government regulations on energy use. These adjustments in input patterns involve reduced intensity of energy use, greater intensity of labor input, less use of nonenergy intermediate materials in most sectors, and, apart from services where capital input increases, a reduction in the relative importance of capital services. Final demand patterns alter in response to the policy measures, partly as a result of the changing price structure and partly as a result of government regulations on energy use patterns. Table 2 also shows the effects of energy policy in final demand patterns for 2000. As a result of final demand patterns and the structure of inputs into production both shifting away from energy-intensive goods and services, the pattern of gross output also shifts away from energy; these changes are also summarized in Table 2.

III. Economic Growth

Finally, we consider the effects on economic growth of restrictions on energy consumption. Table 3 summarizes the aggregate economic effects of the four policies in the year 2000. The level of real *GNP* in 2000 is reduced, relative to the base case, by 1.5 percent for Policy 1, 3.2 percent for Policy 2, 7.2 percent for Policy 3, and 11.9 percent for Policy 4. Both consumption and investment are reduced; the reduction in consumption is greater than that in investment in Policy 1 but the reverse is true for the other policies, that is, the more stringent the policy, the greater the relative impact on investment. The investment impact is particularly important since it leads to a slowing of the rate of growth of productive capacity and of

TABLE 3—ECONOMIC IMPACT OF ENERGY POLICIES IN THE YEAR 2000

	Base Case	Policy 1	2	3	4
GNP ($1972 billion)	2721.7	2679.8	2634.9	2524.6	2397.0
Percentage difference		−1.5	−3.2	−7.2	−11.9
Average annual growth rate, 1977–2000 (percent)	3.2	3.1	3.0	2.8	2.6
Consumption ($1972 billion)	1763.5	1733.1	1706.4	1622.2	1520.3
Percentage difference		−1.7	−3.2	−8.0	−13.8
Investment ($1972 billion)	401.1	394.7	381.4	359.0	335.8
Percentage difference		−1.6	−4.9	−10.5	−16.3

output. By 2000, capital stock in Policy 4 is 11 percent below the base case level. This reduction in capital input accounts for 3.2 of the 11.9 percent reduction in real *GNP*. In subsequent years the relative importance of this dynamic effect is still larger. We conclude, then, that policies to restrict the growth of energy consumption have the potential to achieve the specified objectives of reduced energy growth, reduction in dependence on imported energy sources, and increased use of relatively abundant domestic energy sources, but that these changes involve possibly large economic cost in terms of slowed economic growth and output foregone. Reduction in energy growth is not a desirable social objective in itself. The benefits assigned to this reduction must be balanced against these costs.

REFERENCES

M. Beller, *Sourcebook for Energy Assessment*, BNL 50483, Upton, New York 1975.

E. R. Berndt and D. W. Jorgenson, "Production Structure," in Dale W. Jorgenson et al., eds., *Energy Resources and Economic Growth*, final report to the Energy Policy Project, Washington, Sept. 1973, ch. 3.

——— **and D. O. Wood,** "Technology, Prices, and the Derived Demand for Energy," *Rev. Econ. Statist.*, Aug. 1975, *57*, 259–68.

——— **and** ———, "Engineering and Econometric Approaches to Industrial Energy Conservation and Capital Formation: A Reconciliation," Mass. Inst. Technology Energy Lab., work. paper no. MIT-EL-77-040WP, Nov. 1977.

C. W. Bullard and R. A. Herendeen, (1973a) *Energy Cost of Consumption Decisions*, document 135, Center for Advanced Computation, Univ. Ill. 1973.

——— **and** ———, (1973b) *Energy Cost of Consumer Goods 1963/67*, document 140, Center for Advanced Computation, Univ. Ill. 1973.

E. A. Cherniavsky, *Linear Programming and Technology Assessment*, BNL 20053, Upton, New York 1975.

L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Production Frontiers," *Rev. Econ. Statist.*, Feb. 1973, *55*, 28–45.

———, ———, **and** ———, "Transcendental Logarithmic Utility Functions," *Amer. Econ. Rev.*, June 1975, *65*, 367–83.

K. C. Hoffman, "A Unified Framework for Energy System Planning," in M. F. Searl, ed., *Energy Modeling*, Washington 1973.

——— **and E. A. Cherniavsky,** *Interfuel Substitution and Technological Change*, BNL 18919, Upton, New York 1974.

——— **and D. W. Jorgenson,** "Economic and Technological Models for Evaluation of Energy Policy," *Bell J. Econ.*, Autumn 1977, *8*, 444–66.

W. W. Hogan, (1977a) "Capital Energy Complementarity in Aggregate Energy-Economic Analysis," unpublished paper, Energy Modeling Forum, Inst. Energy Stud., Stanford Univ., Sept. 1977.

———, (1977b) "Energy and the Econ-

omy," unpublished paper, Energy Modeling Forum, Inst. Energy Stud., Stanford Univ., Sept. 1977.

E. A. Hudson and D. W. Jorgenson, "Interindustry Transactions," in Dale W. Jorgenson et al., eds., *Energy Resources and Economic Growth*, final report to the Energy Policy Project, Washington, Sept. 1973, ch. 5.

——— **and** ———, "U.S. Energy Policy and Economic Growth, 1975-2000," *Bell J. Econ.*, Autumn 1974, *5*, 461-514.

——— **and** ———, "The Long Term Interindustry Transactions Model: A Simulation Model for Energy and Economic Analysis," final report to the Applied Economics Division, Federal Preparedness Agency, General Services Administration, Washington, Sept. 1977.

Jack Faucett Associates, *Data Development for the Input-Output Energy Model*, final report to the Energy Policy Project, Washington 1973.

D. W. Jorgenson, "Consumer Demand for Energy," in William D. Nordhaus, ed., *Proceedings of the Workshop on Energy Demand*, Laxenburg 1975, 765-802.

——— **and L. J. Lau,** "The Structure of Consumer Preferences," *Annals Soc. Econ. Measure.*, Jan. 1975, *4*, 49-101.

L. Kantorovich and T. C. Koopmans, "Problems of Application of Optimization Methods in Industry," Federation of Swedish Industries, Stockholm, Nov. 1976.

W. Marcuse et al., *A Dynamic Time Dependent Model for the Analysis of Alternative Energy Policies*, BNL 19406, Upton, New York 1975.

DISCUSSION

TJALLING C. KOOPMANS, Yale University: The study presented by Edward Hudson and Dale Jorgenson rests on an impressive body of work in the econometric representation of production possibilities developed over the last twenty years or so. Two different methods are applied: one to the energy sector; the other to the industries that make up the rest of the economy. While their description of the latter method uses the terminology of the classical input-output approach developed by Wassily Leontief, the method is essentially different in that it both recognizes and estimates the effects of the relative prices of the various inputs on the input mix used in each industry. This is achieved by econometric estimation of the price possibility frontier of each nonenergy industry from past data. (This is a relationship between prices that, under the assumption of profit maximization in competitive markets, is dual to, is implied in, and in turn implies the production function.)

The energy sector is represented by a process model developed by Hoffman, Marcuse, and their colleagues at Brookhaven Laboratory, based on engineering estimates of ratios of inputs to outputs for a substantial number of energy extraction, transportation, conversion, and utilization processes. The essential point here is that the number of processes in the model exceeds the number of different end uses, thus enabling the model to simulate a cost minimizing or benefit-minus-cost maximizing response to demand shifts or input price changes.

The principal advantage of the process model over the "econometric" model used for the nonenergy industries is that it can absorb not only engineering information based on past performance, but also anticipated characteristics of processes not yet in operation, at all, or on a commercial scale. Moreover, the process model allows one to simulate the effects of quantitative or other constraints placed on the levels of operation of specific processes for reasons of environmental protection.

The greater flexibility of the process model raises the question why it is not developed for some nonenergy industries as well. As Jorgenson has explained, the more flexible model is also by far the more expensive one, data-wise. I suggest that just the same the possible extension of the process model to other industries is to be guided by cost-benefit analyses.

In conclusion, I have a few questions to the authors:

1) Could a test of the process model and the (primal or dual) production function model against each other be made for the energy sector over a past period where the data needed for both methods are available?

2) For the econometric estimates of production relations, could standard errors of estimate be reported?

3) It is my understanding that the surprising energy-capital complementarity conclusion is based not only on the estimated (primal or dual) production relations, but also on reasoning that includes additional relations of economic behavior. Have these been specified, and where?

4) Can an aggregate "price elasticity of demand for energy" be read off from the model solutions? If so, what is its value, with the price being measured at what point in the chain from extraction to end use?

5) Finally, a question related to the gauge by which the impacts of policies constraining energy use are measured. If that gauge is the discounted sum of future utilities from total consumption, then could the effect of the energy-capital complementarity be a shifting of consumption to a nearer future, leaving perhaps only a second-order effect of uncertain sign on that gauge?

CLARK W. BULLARD, U.S. Department of Energy: The physical scientist views the flow of energy through a system in a fundamentally different way than the economist. My remarks are intended to highlight those differences, as well as to assess the role of economic modeling in the analysis of long range energy policy issues.

The physical scientist views consumers as creatures who "demand" changes in their sensory inputs. They meet these demands by changing the thermodynamic state of the earth's resources around them, or by transporting themselves to a different environment. To change material resources to a state different from their natural condition requires energy, and the second law of thermodynamics places a lower limit on the amount of energy required. What the economist calls capital equipment is seen by the physicist as a conduit for channeling fuel energy to do useful work. For millenia, solar energy was channeled through land and thence through domesticated animals, slaves, or independent laborers to provide the dominant energy source for economic activity. During the last century, fuel energy grew to comprise over 90 percent of the total, and the ideas of Ricardo have given way to new modeling approaches.

Hudson and Jorgenson's overall framework for modeling energy flow through the economic system represents a substantial improvement over earlier attempts. In theory, the interindustry submodel allows for explicit consideration of physical limits to energy efficiency, and the consumption model for saturation of energy-consuming activities such as air conditioning and time spent driving. Due to aggregation, however, some factors to which energy demand is highly sensitive cannot be adequately represented in the present versions of the model. The optimization submodel is quite detailed, and reflects the explicit recognition that characteristics of energy-producing technologies of the year 2000 are known, including their potential contribution to growth in total factor productivity. Had the energy-consuming sectors been modeled in similar detail, it would have been unnecessary to assume a relatively crude exogenous extrapolation of past trends for productivity growth. I shall return to this issue below, after considering the empirical basis for the representation of interindustry energy flows.

Hudson and Jorgenson's result hangs on the assumption that the relations among energy, capital, and labor during the technological changes of the postwar period will hold during the coming decades. In recent work, Ernst Berndt and D. Wood have observed that the data indicate that production technologies became less labor intensive while the relation between energy and capital inputs remained relatively stable. This would be consistent with an argument that industrial *R&D* was intended to replace labor by relatively dependable machines powered by electricity and gas, and that advances in control technology reduced the labor intensity of production. Since postwar energy costs were a small and decreasing fraction of almost everyone's budget, industrial designers generally paid little attention to the potential for substituting capital for energy. Recent events, however, have shown American industry to be as vulnerable to interruption of its price-regulated gas supply and oil-generated electricity supply as to its strike-prone labor supply. It seems reasonable, therefore, to expect engineers to substitute capital for energy in modern industrial facilities. Such a trend could reduce the economic losses forecast by Hudson and Jorgenson to result from higher energy prices.

My second area of concern relates to the policy relevance of the chosen base case. Most earlier analyses of this type have assumed the relative world price of oil remained constant; Hudson and Jorgenson's analysis breaks new ground in assuming a 1 percent per annum increase. Recent analyses of the world oil situation, however, forecast conditions during the mid-1980's calling for a significantly higher price assumption. Closer to home, the congressional debate over the president's energy proposal has the House supporting Scenario 1, and the Senate something more like Scenario 2. Apparently, few people believe the full-employment, maximum *GNP* growth trajectory of the "base case" is likely to occur. Its use as a yardstick for measuring the impact of an energy tax policy in Scenario 3 is, therefore, questionable.

It is useful to point out that low energy growth scenarios such as Policies 2 and 3

are not necessarily austere. The National Academy of Sciences Committee on Nuclear and Alternative Energy Systems recently examined several such scenarios, and from an energy pricing perspective two of them bear a striking resemblance to Policies 2 and 3. I can relate some impressions gained during the two years that I chaired the Demand Analysis Integration Resource Group. Scenario 3 could find us spending 40 percent more time in efficient 30 mpg cars almost as spacious and comfortable as today's. Per capita air travel would be up 60 percent and we would be living in better-built air-conditioned homes that use less than half as much energy as today's. The quality of the environment would be improved dramatically by eliminating impacts associated with producing and consuming about 50 quadrillion fewer Btu of energy, compared to Hudson and Jorgenson's base case. (By comparison, today's domestic production is about 60 quadrillion Btu.)

As William Hogan and A. Manne have demonstrated, *GNP* growth is extremely sensitive to the long-run energy price elasticity. The structural details added to this energy model have undoubtedly improved the estimate of this variable and extended the model's useful time horizon. But in using it to analyze human behavior at the turn of the century, we must remember that nearly half the population of that year has not yet entered kindergarten; their values and price elasticities of energy demand may be substantially different from our own. Due to this fact and our lack of experience with energy prices several times today's level, the empirical basis for the price elasticities used in Scenarios 3 and 4 is extremely questionable. Finally, the relative prices lying at the heart of the model are critically dependent on the growth of total factor productivity, the largest single contribution to *GNP* growth in the model, which is assumed to grow at some exogenous rate extrapolated from historical trends. Beyond some time horizon, the secular variations in this parameter cannot be forecast; they are by definition unknowable.

Faced with the inherent limitations of economic models to characterize the world beyond the turn of the century, what factors should guide our energy policy? We find ourselves considering technologies having ten to fifteen-year construction times and thirty-year economic lifetimes, asking for reasonable assurances that the technologies will be economic and legal to operate during that period. Energy developers cry for Congress to provide certainty in the regulatory environment; yet the United States Constitution prevents one Congress from binding the next, and guarantees Americans the liberty to throw the rascals out every two, four, or six years to reflect fundamental changes in our society's values. Certainly flexibility must be a key factor to consider in our energy investment decisions.

Another factor of growing importance is environmental quality. Rising imports have masked the fact that growth in domestic energy production has been virtually stopped since before Earth Day.

A growing number of people believe society may be threatened more by too much energy too soon than by too little too late. John Holdren has put it succinctly: "The total amount of energy a society can have in the long run will be limited . . . not by resource availability or by economics narrowly defined, but rather by the rising environmental and social costs of energy supply in the future." The useful time horizon of energy models is limited by the rate of evolution of such public attitudes.

Finally, future energy policy must be guided by an increased respect on the part of Americans for the rights of other individuals and nations to behave occasionally in a manner that may seem to us "economically irrational." It is especially unfortunate that many conventional economic modeling activites, particularly those focusing on *OPEC* as a market defect, do little to foster such tolerance. Still savoring the lingering taste of Manifest Destiny, it may be difficult for we newcomers to the North American continent to understand how the local presence of a blood heritage dating back thousands of years might

endow one with a discount function much more complex than the simple exponential parameter in an economic model. And until we feel responsible to the rest of the world for imposing quotas on the production of food while one-fourth the world's people are undernourished, I doubt many will pay attention to our cries from grounded gas guzzlers for other sovereign nations to accelerate depletion of their petroleum reserves at a rate dictated by our economic models.

Recalling the physicist's definition of power as the rate of energy flow per unit time, it becomes apparent that energy investment decisions will be based on equity at least as much as on efficiency. Power, in this sense, is directed by the owners of conduits through which energy flows, and by those capable of interrupting its flow. As the ongoing congressional debate highlights these equity concerns, we can only feel less comfortable with models that assume the issue away.

WILLIAM W. HOGAN, Stanford University: As is their custom, Edward Hudson and Dale Jorgenson have compressed many issues into a few carefully chosen words. The policy tests they present benefit from a substantial heritage of development of models of the energy sector in the full economy. Their qualitative results and, in an approximate way, their quantitative estimates represent or extend the state of the art of the analysis of energy and economic growth. The boundaries of these analyses are expanding rapidly, at least in terms of the understanding of the application of theory, and it may be useful to highlight some of the relevant issues by way of comparison to other studies and other models. These remarks draw extensively on the recent examination of six models of energy and the economy conducted by the Energy Modeling Forum (*EMF*).

The implicit model behind many energy policy analyses is an assumption of a linear relationship between *GNP* and energy. This naive view was challenged most emphatically in the Ford-Mitre study which focused on the small value share of the energy sector to conclude that a very weak link existed in the long run and that energy growth could be curtailed without a similar reduction in *GNP*. After accounting for the possibility of less than unitary energy demand price elasticities, this qualitative conclusion was confirmed in the application of several energy sector partial equilibrium models by the Modeling Resources Group of the Committee on Nuclear and Alternative Energy Systems. Compared to the naive model, this qualitative conclusion of a weak link between energy and the *GNP* persists in the results of Hudson and Jorgenson's general equilibrium model.

The many implications of this result, and its counterintuitive nature for many policymakers, lead the *EMF* to investigate the same question through the comparison of several general equilibrium models of energy and the economy. This *EMF* study confirmed the results of the partial equilibrium models on the importance of the small value share of energy in the economy and further highlighted the sensitivity of the feedback effect to the magnitude of the elasticity of substitution between energy and other primary inputs. The results of Hudson and Jorgenson are representative of those obtained by the *EMF* for models which include empirical estimates of the substitution potential. The main new contribution of the general equilibrium models is the identification of important dynamic interactions, primarily through the effects of energy scarcity on aggregate capital investment. Increased energy prices reduce the demand for energy and also reduce the productivity of capital. With a lower rate of return, investment decreases and the long-run capacity and output of the economy are reduced. This indirect effect on capital formation works in the same direction as the energy scarcity and has an effect on long-run *GNP* of the same order of magnitude as the direct impact of the energy reduction.

It is easy to be complacent about these measures of small economic impact. For example, Table 3 in the Hudson and Jorgenson paper shows that a decline in energy

consumption is possible with a substantial increase in aggregate real output. Hudson and Jorgenson, however, point out that the loss in real output is large in absolute terms and the benefits of lowered oil imports must be balanced against these costs. Of course, the difference in the year 2000 of a few hundred billions of dollars against a few trillion dollars may be hard to grasp. In this regard, the authors have done us a service by providing the estimates of the prices required to achieve the parsimonious use of energy. I find the $7.00/barrel tax of Policy 2 unlikely without some outside assistance, much less the $50.00/barrel tax of Policy 4. In addition, the richness of the model properly illustrates that the impacts of energy taxes will not be evenly spread over all sectors, nor, I would guess, over all income classes or regions of the country. The small proportional effects on *GNP*, therefore, may deflate the naive argument about the link between energy and the economy, but that should not make us too sanguine about the role of energy.

This is not the complete story either. The model examined here and those of the *EMF* study focus on the long-run potential of the economy. Abrupt changes in energy availability or other policies with short-term implications may affect the realization of this potential *GNP*, but they are not within the scope of the models. Further the models require assumptions about the future population or labor force growth and the rate of technological change which, other things being equal, determine the growth path of the *GNP*. There may be some effect of energy price or energy availability on the variables whose values are here assumed. Any such effects would not be captured in the models. In addition, the effects of regulation, industrial organization, environmental considerations, and interactions with the financial sector are excluded or addressed in a rudimentary manner. The models have come a long way from the naive hypothesis of the linear relationship between *GNP* and energy, and there are many important questions that can be answered, at least in part, but there is ample room for more work. The clear priorities are for the analysis of the short-term adjustments and the compositional effects of changes in energy policy.

LESTER B. LAVE, Carnegie-Mellon University: Next to teaching undergraduates, energy modeling is the fastest growing indoor sport for economists. Concern for the future has a high income elasticity; naturally, we Americans, particularly middle-class intellectuals, seem to focus much concern on the twenty-first century.

It is all too apparent that our destiny results from the clash of such giants as future rates of population growth, technological change, and such inherently unpredictable events as wars, plagues, and droughts. We have little ability to modify these forces that shape the future, although the forces can be changed at the margin and we can take actions that take advantage of fortuitous events and protect against adverse ones. So, along with watching the leviathans lash about, upsetting our lives, we study the future, plan, and worry.

To get some perspective on the controversy, I note that most economists are incurable optimists, despite Malthus. The implication of virtually every energy modeling exercise by economists is that per capita income will continue to grow within a range of 1–2 percent per year for the next half century. If true, economists are assuming that the growth in real income might slow somewhat, but it will still continue to be true that our grandchildren will be much richer than we are.

In four years, we have meandered through three stages of conclusions regarding the effect of energy on *GNP*. In the first stage, energy and *GNP* were believed to be bound together in essentially a one-to-one relationship. The United States' historical experience from 1950 to 1970 and a casual look at the energy consumption of rich versus poor nations seems to imply this direct link.

This evidence seemed all the more unfortunate in view of the growing opposition to the Alaska Pipeline, nuclear reactors, the breeder reactor, and other types of electricity generation facilities in particular and to energy projects in general. Investment in

energy facilities seems to have slowed, promising future energy shortages and slower economic growth in the future. As a result warnings have been sounded about the need to get on with building energy facilities, especially since many of the industry spokesmen regard current facilities as safe and not destructive of the environment. In these views we must clear away impediments and speed the development of domestic energy supply. This is the world with *U.S.* consumption of 300 quads in 2010 and 1200 quads in 2050.

If the production function for *GNP* were of a fixed coefficient form, we would face tough decisions about trading off environmental quality and safety for economic growth. We might want to subsidize energy use so that we could promote its growth in order to promote economic growth. A few billion dollars in an energy subsidy would serve to increase *GNP* by a few hundred billion dollars.

Fortunately, the elasticity of substitution of energy for other inputs is not zero. Empirical work suggested elasticities of substitution as close to unity as to zero. Thus the second stage of energy modeling incorporates a nonzero elasticity of substitution.

In a simple model with a single sector other than energy, an elasticity of substitution of .5 or more implies that energy use could be reduced one-half to two-thirds with a resulting reduction in *GNP* of only a few percentage points (given time to adjust the capital stock). Not so for an elasticity of substitution of .2 or less. There an attempt to reduce energy use by one-half would result in a large fall in *GNP*.

Seemingly, a high elasticity of substitution would give us our cake and let us eat it too. We could have economic growth along with diminished energy use. The second stage focuses attention on the elasticity of substitution. Unfortunately, the story gets more complicated since the economy is not a two-sector model; to estimate the disaggregate elasticities of substitution, one has to evaluate complicated econometric techniques and interpret sectoral estimates. However, the predominant conclusion is still hopeful.

Enter the revisionists of the third stage to protest that it is not so simple. Since capital and labor are substitutes and since each is presumably a substitute for energy, increasing the price of energy will drive up its marginal product and drive down the marginal products of capital and labor. Presumably, the brunt of the reduction will be borne by the return on capital. If so, capital formulation would fall and eventually the growth rate of *GNP* would slow. Thus, it appears that capital and energy might be complements in practice, not substitutes, a result frequently found in empirical work. Therefore the effect of a slower growth in energy use would be a slower rate of capital formation, and eventually, a slower growth rate of *GNP*—in spite of a high elasticity of substitution.

A supporting argument is that productivity growth differs across sectors, with services slowest and manufacturing fastest. Surely, energy shortages, combined with higher prices for energy, would lead to shifts in both production and consumption patterns; there would be a gradual shift from the high productivity, high energy using sectors to the low productivity, low energy using sectors (from manufacturing to services). If so, growth would slow, even independently of the expected slowing of capital formation.

Both of these arguments destroy, or at least mitigate, the optimism of the second stage. But there is a puzzle here. These two arguments seem to imply that the second-order effects will be negative, mitigating the substitution effect. It is puzzling that both these effects seem to speed substitution, but are cited as factors slowing *GNP* growth. Surely, the shifts to a less energy intensive *GNP* are a benefit, compared with constraining adjustment so that *GNP* was produced in the old way by subsidizing energy use. I suspect that one difficulty is in the way *GNP* is measured. If the Nordhaus-Tobin Measurement of Economic Welfare or some similar measure of welfare were used, these secondary adjustments would be shown to increase, not decrease, the growth in economic welfare.

A more basic difficulty with the third stage arguments is that many of the crucial variables are endogenous. In particular the

return to capital is largely within government control. The tax treatment of depreciation, investment expenses, and business profits can change the rate of return from negative to positive or vice versa. Savings and capital formation can be increased by government assuming the burden of collecting taxes to force savings (for example, social security as a vested pension plan), strengthening incentives for private savings and changing tax laws to make investment more attractive.

Thus, if society thinks that saving and investment are too small, a few twists of the policy instruments can cure that. Furthermore, we could change production techniques and the sectoral composition of demand by tax policy, if we should desire. Finally, *R&D* can be targeted to increase the elasticity of substitution.

In conclusion, the lesson is that we can have our cake and eat it too, if only we knew what cake we wanted! The growth of energy use can be hastened, slowed, or left alone; capital formation and sector shares can be changed. The important questions relate to our goals. Do we want massive industrial output? Vast changes in life style? Orwell's *1984* or Schumacher's *Small is Beautiful?*

Thus the analysis ends by asking ourselves those questions that economists dislike facing. We middle-class intellectuals must ask how we want to live and work in the twenty-first century. If goals can be agreed upon, I forecast that economists could get busy modeling the policy instruments and analyzing how each will affect the future. We would examine the primary effects to see how to achieve our goals, and then the secondary and tertiary effects to ensure that we reach them.

Unfortunately, as with so much of economic policy, we must define our goals before we can decide which current policies are good, or in what ways we ought to change them. As recent energy debates have made apparent, neither in public nor in private is there general agreement about where we should be going.

Disembodied Technical Progress: Does Employee Participation In Decision Making Contribute To Change and Growth?

By KARL-OLOF FAXÉN*

As part of a joint union-management research project on the effects of employee participation, detailed studies of changes in work organization and the development of productivity have been conducted in a number of member firms of the Swedish Employers' Confederation. This paper reports the results of these studies and offers an interpretation of the effects of employee participation on disembodied technical progress at the plant level.

Disembodied technical progress constitutes the major part of productivity increase in low capital intensity branches of manufacturing as well as in building. In process-type, capital-intensive branches of manufacturing, it is insignificant (see Yngve Åberg, 1969). Disembodied technical progress accounted for one-third of productivity gains in Swedish manufacturing 1951–69 and increased to an annual rate of almost 3 percent, 1966–69 (see Åberg, 1971).

The entire program of studies covered ten companies in different sectors—engineering (Sickla and Åkers), chemicals (Perstorp), insurance (Skandia), etc. The interaction between organizational change and the psychological reactions among workers, supervisors, technicians of various categories, etc., were studied along with productivity. Positive as well as negative conditions for organizational change and their relation to disembodied technical progress were identified.

*Swedish Employers' Confederation.

I. Perstorp—Technical Laminates

Perstorp is a chemical plant with a diversified production program. The experiment as described in Lars Forsberg et al. took place in the technical laminates division with about sixty workers. This division was composed of three departments, Spraying, Pressing, and Finishing. Capital intensity and machine dependence were lowest in Finishing; Spraying had a process-type, capital-intensive technology, while Pressing was mixed.

After a year of preparation, the experiment started in January 1971 and lasted until the end of February 1972. A number of changes were introduced after agreement by the majority of participants. Day-to-day decisions on production matters were transferred from supervisors to workers. Increased cooperation was emphasized.

A special analysis in March 1971, three months after the start of the experiment, indicated that conditions for cooperation were most favorable in Finishing and least favorable in Pressing. This difference in initial conditions was reflected in the development of productivity: a 23 percent annual increase 1970–71 in Finishing versus a negligible gain in Spraying and a 5 percent decrease in Pressing. In rankings of the departments by workers' subjective perceptions of changes in job content, Finishing was again foremost.

The process of change in Finishing involved greater consultation among workers in special situations, in order to avoid rejects by rescheduling machine operations. As worker confidence grew, they sought

specialist technical advice more frequently, and their positive experiences stimulated continued experimentation. It was easy to observe the effects on productivity directly on the shop floor. The quality of working life improved.

Similar processes of change could not take hold so readily in the other two departments. Some initiatives were noted, but they did not result in much change. This was not solely due to technology. In these cases the individual workers' tasks were not interdependent but ran in parallel, making it more difficult to achieve concurrent productivity gains. It was not easy to discern changes in productivity in an informal way; they required statistical measurement. The quality of working life did not improve significantly.

These results indicate productivity can be increased substantially if organization and working methods are improved as a result of employee participation. In the following, some general conclusions from the Perstorp and the other experiments will be indicated.

II. Technical Learning, Social Learning, and Feedback of Information

Organizational development is a process through which the overall competence of the organization is enhanced. Its operating and managerial methods improve, along with its capacity to solve problems and adjust to changes in its environment. Increases in productivity achieved via organizational development—disembodied technical progress at plant level—presuppose learning of two kinds, technical and social.

Technical learning can take the form of a worker learning to handle more machines, a supervisor learning how to run a new planning system, or a production technician learning new methods of work measurement. Social learning involves learning about other people and groups in the work organization, about their motives, reactions, values and ambitions. It also relates to the way in which one perceives oneself and one's role in the organization in such matters as responsibility and authority, operating methods and specific expertise.

Learning is related to changes in the day-to-day work situation. A new way of doing one's own work often means changes in social relationships with other people in the workplace. New ways of working bring new experiences, both technical and social. When problems are encountered in the ongoing work and in the relationships with other people at work, they constitute a new challenge. Learning is experienced as problem solving.

Learning, whether social or technical, requires feedback about the results of changes. Positive feedback reinforces a continuation of learning. Without positive feedback there will be little learning. Positive feedback makes people feel learning is worthwhile.

Feedback information can take many forms: for example, statistics, diagrams, gestures or facial expressions. The structure of the work organization controls the possibilities for various types of feedback. Formal information can be transmitted via the channels indicated in the organizations plan, proceeding from one department to another, or from a higher to a lower level, and vice versa. These formal flows are the easiest to observe; one can see whether they contain positive or negative feedback. However, one must not imagine that all the feedback flows through the formal organization—informal contracts are also important.

If the feedback of results is to be accepted, people in an organization must trust one another. If groups or individuals distrust the productivity measure, and think it is misleading or meaningless, learning cannot develop. This kind of attitude is stiffened if the changes constitute a threat to the way in which these groups interpret their roles. At Perstorp it was common for people who felt threatened by the changes in organization and methods to distrust the productivity measures.

In sum, feedback information will not be accepted if the framework is a social rela-

tionship that is perceived negatively. Social learning must then occur if technical learning is to be reinforced. Attitudes to other employees have to change, and so has one's interpretation of one's own role at work. For example, a technical problem may be overcome, leading to a new operating method, and altering social relationships. Provided the results of the change are transmitted back to and are interpreted in a positive way by the people affected, attitudes can change and people can interpret their roles in a new way. Social learning commences. Technical information can be received and processed in a still more positive manner. The new operating method is learned in an active way. A successful new method thus involves an interplay between technical and social learning.

Social learning resides in the organization and is transferred to new employees. A new employee learns from more experienced workmates how to behave to the supervisor, production technicians and other groups with whom he comes in contact. Social learning can be expected to show greater survival power than technical learning. Every one of the workers at Sickla learned additional technical aspects of the assembly work in the course of the experiment. That was of course excellent in itself, but it was not particularly helpful when later new and heavier types of machines were to be assembled. Social learning, in the form of improved relations between workers and production technicians, was however still present. As a result a new type of joint endeavor to find the best methods of assembling the heavy machines developed.

III. Rewards

A better understanding of both technical and social relationships rewards problem solving in a way that does not exhaust itself over time. Also, it does not conflict with employee participation. Improvements in quality of working life were primarily seen in that frame of reference. They were observed by means of periodical interviews structured so as to permit analysis of how work becomes more meaningful as the individual improves his ability to relate his work to wider goals.

Pay structure and the formal and informal status rankings between different jobs were found to be important parts of the reward system in all experiments. Changes in work organization required changes in pay systems as established in agreements with trade union locals. At Skandia changes were made in the form and method of payment of qualification premia. Sickla changed over to a group bonus wage, Perstorp to a straight-time rate. Åkers also adopted a group bonus scheme. Thus, the formal reward structure was adjusted. As was particularly evident in the Skandia experiment, the informal status ranking between jobs did resist change. This prevented reinforcement of the social and technical learning which would have made possible less functional specialization in the work organization.

IV. Testing New Methods, and Confidence in the Productivity Measures

Increases in productivity take different forms: more units per hour worked, better quality, less consumption of tools, etc. Correct measurement often requires sophisticated index methods. Productivity measures are frequently deficient, making it difficult to feed back the results of a new method. In the Perstorp Finishing department informal day-to-day productivity measures played the major role in the learning process. The official monthly index numbers were too slow. In Skandia, informal measures were also used for several years until they were found misleading. This event constituted a major breakthrough in the learning process. In Sickla, the productivity measure was only partial and the major part of the increase was never reported to management. In Åkers, groups did not agree on the basis for comparison.

Unsuccessful attempts to apply a new method can be just as significant for learn-

ing as successful ones. The important thing is that people trust the productivity measure and are able to realize whether the experiment is a success or a failure.

At Sickla, it was important for the development of the learning process that the workers were permitted to test an alternative method of assembly during a two-week period. A large table was used instead of the regular assembly line. Contrary to the predictions of the production technicians and others, productivity was maintained, and this fact stimulated general interest in the continuation of the experiment. The workers learned more about technical problems of assembly and how to make more flexible use of the regular assembly equipment. Their attitude towards it became more positive. They also decided not to continue working at the table because they found it too exhausting.

V. The Quality of Working Life and Work Intensity Under Participation

One important conclusion from the experiments, which emerges most clearly in the case of Skandia, is the individual character of the quality of working life. The prospect of new and more extensive duties may be a path to greater involvement in his work for one person; while for another, the same prospect may mean greater demands upon him, which must be resisted.

Even when representatives of the involved workers and staff participate in decision making, changes can have adverse consequences for some. Employee participation does not guarantee that everyone will experience positive changes in quality of working life. The workers at Perstorp discussed minority rights extensively. On their initiative, a qualified majority rule was adopted for major decisions as a means of protecting minorities.

It is not possible to formulate simple and generally valid rules as to the way in which an organizational change should be designed to minimize adverse effects on some person or persons. Of course, this does mean that participation would produce worse decisions in this respect than more traditional forms of decision making.

It is sometimes said that the intensity of work is affected by people having an opportunity to take part in decision making. The employees on the shop floor allegedly feel more involved in and responsible for decisions which they have helped to make, and therefore, feel obliged to work harder and use their working time to better effect. The experiments do not support this view.

VI. Conclusion

The productivity increases found in the experiments depended both on improved coordination and on the fact that joint problem solving led to improved operating methods. Less functional specialization was made possible by enhanced technical learning. Participation facilitated social learning. This in turn removed obstacles for technical learning. Observed productivity increases did not depend on increases in work intensity or exertion.

None of the experiments demonstrated a simultaneous improvement in the quality of working life and a decline in productivity. Where productivity did not progress in a satisfactory way, there appears to have been some associated conflict or uncertainty which also found expression in an unsatisfactory trend in the quality of working life.

REFERENCES

Lars Forsberg, Reine Hansson, and Jan Pärsson, "Försök med ändrad arbetsutformning och arbetsorganisation på en avdelning av Perstorp AB," Delrapport 1–4, Stockholm 1974.

Yngve Åberg, *Produktion och produktivitet i Sverige 1861–1965*, Uppsala 1969.

———, "Långtidsutredningen och teknikfaktorn," Industriförbundets Tidskrift, No. 5, 1971, 15–19.

Job Satisfaction as an Economic Variable

By R. B. FREEMAN*

Job satisfaction, while the subject of popular attention, of an extensive sociology and industrial psychology literature, and of theories of "alienation," has been studied by relatively few economists (see George Borjas; Daniel Hamermesh; Robert Flanagan, George Straus, and Lloyd Ulman). Partly, the neglect of job satisfaction reflects professional suspicion of what may be called *subjective variables:* variables that measure "what people say" rather than "what people do." Partly also, economists are leary of what purport to be measures of individual utility.

The purpose of this paper is to examine these concerns and evaluate the use of job satisfaction (and other subjective variables) in labor market analysis. The main theme is that, while there are good reasons to treat subjective variables gingerly, the answers to questions about how people feel toward their job are not meaningless but rather convey useful information about economic life that should not be ignored. The paper begins with a brief description of the satisfaction questions on major worker surveys and then considers the use of satisfaction as an independent and as a dependent variable. Satisfaction is shown to be a major determinant of labor market mobility, in part, it is argued, because it reflects aspects of the work place not captured by standard objective variables. Satisfaction is also found to depend anomolously on some economic variables (such as unionism) in ways that provide insight into how those factors affect people.

*Harvard University. I have benefited from discussions with Robert Fogel, James Medoff (who suggested the Z-score transformation), Larry Summers, and from the research assistance of Laura Nelson and Eric Seiler.

I. The Job Satisfaction Variable

To begin with, Table 1 reproduces the job satisfaction questions and distributions of responses from major surveys of workers. The satisfaction questions are quite similar across surveys, asking for an overall evaluation of job satisfaction, and invoked similar distributions of responses. Most persons report themselves as highly or quite satisfied with their jobs, with only a distinct minority of about 10 percent reporting dissatisfaction. While there is some indication in the National Longitudinal Survey *(NLS)* longitudinal tapes of declines in satisfaction over time, the Michigan Work Quality and Quality of Employment Surveys shows no such pattern.

The responses to satisfaction questions (and other subjective variables that lack a definite metric) can be scaled in two possible ways in analysis. First, they can be written as n-chotomous variables, taking the value 1 if the individual's response fell into the given category and 0 otherwise. When satisfaction is an independent variable, the set of dummies has an a priori ordering of effects with, for example, the third category having a larger effect than the second (relative to, say, the first) and the fourth a larger effect than the third. When satisfaction is the dependent variable, the multinomial probability model can be used to predict the effect of various factors on the probability of giving a certain response. Alternatively, the variable can be rescaled according to a specified symmetric probability distribution, such as the standard normal. With the unit normal transformation, satisfaction becomes a Z-score measuring the number of standard deviations between a given response and the mean. This procedure yields a continuous variable that can be entered as a dependent or independent factor in linear regressions,

TABLE 1—QUESTIONS ABOUT JOB SATISFACTION AND RESPONSES TO QUESTIONS FROM MAJOR SURVEYS
(Shown in Percent)

Survey and Year	Question and Response				
National Longitudinal Survey (*NLS*)	"How do you feel about the job you have now?"				
	dislike it very much	dislike it somewhat	like it fairly well	like it very much	
Older Men, 1966	2	5	37	56	
1971	2	6	45	48	
Young Men, 1966	3	8	42	47	
1966	2	9	50	38	
Michigan Work Quality (1968–69) and Quality of Employment (1972–73)	"All in all how satisfied would you say you are with your job?"				
	not at all	not too satisfied	somewhat satisfied	very satisfied	
1968–69	3	11	39	46	
1972–73	2	8	38	52	
Michigan Panel Survey of Income Dynamics (*PSID*)	"In general would you say your job is: . . ."				
	not enjoyable at all	not very enjoyable	somewhat enjoyable	mostly enjoyable	very enjoyable
1972	2	2	21	42	28

Source: Calculated from distribution of answers for the population given by each of the surveys.

with obvious computational advantages over a maximum likelihood multinomial analysis, and will be followed in ensuing empirical work.

II. Behavioral Consequences of Job Satisfaction

Do subjective responses to job satisfaction questions contribute to explaining objective economic behavior? If they do, a case can be made for including subjective variables in analyses of economic activity. If they don't, subjective variables can be safely ignored.

To determine the relation between job satisfaction and overt behavior, the effect of job satisfaction on the behavior most likely to be affected by it, quits, has been estimated using the *NLS* and Michigan Panel Survey of Income Dynamics (*PSID*) longitudinal data tapes. These tapes have the advantage of linking satisfaction in one year to *future* mobility, providing a fix on lines of causality and on the *predictive* power of the variable that is not possible with cross-section data. The impact of satisfaction and other determinants of mobility is studied in terms of a logistic probability function, linking the probability P of quitting a job between years t and s to the characteristics of the person and their initial job in $t(X_{it})$, including job satisfaction:

$$P(Q) = 1/(1 - \exp \Sigma B_i X_{it}) \tag{1}$$

The X variables include standard measures of the objective position of the worker (age, race, sex, education, wage, occupation in the initial job) and ignore for simplicity (and to avoid simultaneity issues) the additional information from the new jobs to which job changers move.

Maximum likelihood estimates of the effect of job satisfaction, measured as a standard normal variable, and of several objective economic factors on quits are given in Table 2, using the logistic form. All of the calculations are limited to wage and salary workers who remained in the labor force in the period considered and who reported all

TABLE 2—MAXIMUM LIKELIHOOD ESTIMATES OF THE COEFFICIENTS AND STANDARD ERRORS OF JOB SATISFACTION AND OTHER VARIABLES ON THE PROBABILITY OF QUITS, USING THE LOGISTIC FORM

Sample Periods and Numbers of Observations	Mean Quits	Logistic Coefficients and Standard Errors: Satisfaction	*ln* Wage	Age	Tenure	Other Variables[a]	Minus *ln* Likelihood
NLS Older Men 1966–71 (3284)	.145	−.31 (.06)	−.37 (.14)	.021 (.013)	−.05 (.006)	2–7, 10	2438
Michigan *PSID* 1972–73 (3730)	.093	−.14 (.06)	−.89 (.12)	−.027 (.006)	−.06 (.01)	1–9, 11	2585
NLS Younger Men 1969–71 (1742)	.123[b]	−.37 (.09)	−.62 (.24)	−.605 (.033)	−.25 (.06)	2–11	596

Source: Calculated from surveys with questions on satisfaction as described in Table 1.

[a]Other variables defined as 1 = sex; 2 = race; 3 = years of schooling; 4 = occupation (7 dummy variables in *NLS* samples; 9 in *PSID*); 5 = industry (9 dummies in *NLS* samples, 5 in *PSID*); 6 = number of dependents; 7 = geographic locale (3 region dummies); 8 = years of work experience; 9 = local market conditions (unemployment in area in *NLS* young men sample; 3 variables reflecting unemployment, shortage of workers, and area wage in *PSID*); 10 = Standard Metropolitan Statistical Areas dummy; 11 = union. Sex, race, and union are dummies.

[b]Quits calculated by a complicated algorithm based on changes from intervening jobs, and is subject to considerable potential error.

the relevant information about their base-year job. Column 1 records the frequency of quits in the three samples. Column 2 records the estimated logistic coefficient for the *Z*-score of satisfaction, scaled so that positive values reflect greater satisfaction; columns 3–5 give the coefficients for *ln* wages, age, and years of tenure with an enterprise. Column 6 lists the other control variables in the calculations, as specified in the table note, while column 7 records the fit of the equation in terms of minus the *ln* of the likelihood function.

The calculations show that, diverse other factors held fixed, the subjective level of job satisfaction is a significant determinant of the probability of quitting, particularly in the *NLS* samples, where it obtains large coefficients four to five times the standard error. The magnitude of the effect of satisfaction on the probability of quitting can be estimated by differentiating the logistic form (1) with respect to the variable, yielding $dP/dX_i = \beta_i P(1 - P)$ which makes the effect of change depend on the level of P. At the mean level of quits, a one standard deviation change in satisfaction changes the probability by .038 in line 1, by .012 in line 2, and by .040 in line 3, all of which are sizeable relative to the means. For comparison, the effect of a standard deviation in the variable most extensively studied by economists, wages, can also be estimated. Multiplying the logistic coefficients in Table 2 by $P(1 - P)$ and the standard deviation of the variable yields the following impact parameters: .024 (line 1), .047 (line 2), and .067 (line 3). By this metric satisfaction has a greater effect than wages on quits in the older male *NLS* data set, a noticeably weaker effect in the *PSID* and a moderately weaker effect in the younger male *NLS* set.

Estimates of the effect of satisfaction on two other measures of mobility; employer initiated separations and total separations, consisting of quits and employer actions, were also made using the same equations as in Table 2. The results showed only slight effects of satisfaction on employer initiated separations (the largest logistic coefficient was −.09 with a standard error of .06 in the older male. *NLS*), but effects on total separations similar to those in the table. By affecting quits, satisfaction alters the overall level of mobility.

While predictive power, statistical significance, and magnitude of effects are not the sole measures of the value of a variable, the evidence on quits in Table 2 does provide a clear answer to the question

with which we began: it shows that subjective expressions of job satisfaction are significantly related to future overt behavior, which makes satisfaction at least potentially analytically useful.

III. Objections and Evaluation

Granting that satisfaction contributes to predicting behavior and is not meaningless, objections can still be raised about its value in social analysis. First, it may be argued that satisfaction is largely a measure of intentions to stay or quit (which could be better captured by a direct "do you intend to quit" question) and thus that the observed impact of the variable simply relates actions to intentions to act, which does not greatly illuminate the causal forces at work. If mobility were the only variable affected by satisfaction or if the effect of satisfaction were eliminated by inclusion of quit intention questions, this objection would have merit. However, the contrary appears true. The industrial psychology literature relates job satisfaction to such forms of behavior as mental health, absences, and physical ailments (see Edwin Locke), suggesting that the variable affects a broader range of phenomena. Inclusion of a direct mobility variable (responses to "what would the wage or salary have to be for you to be willing to take [another job]?" coded 1 if the person responded at no "conceivable pay") barely reduced the coefficient of satisfaction in the *NLS* samples (a drop from .31 to .29 in the older male *NLS*, for example) and contributed less to the explanation of quits than did satisfaction, suggesting that the more general attitudinal variable has greater information content. Inclusion of the variable "have you been thinking about getting a new job?" in the *PSID*, however, did reduce the satisfaction coefficient in line 2 of Table 2 (which was more weakly related to quits than the satisfaction variable in the *NLS*) to insignificance. This would support the objection *if* the intention variable was unrelated to other forms of behavior.

A related deeper problem is that, as a measure of personal feelings, satisfaction may lack systematic independent variation or links to social variables of concern to economists. Assume, for example, that satisfaction depends only on standard measured variables and random noise but does not exhibit any *socially identifiable exogenous* variation. Then it would partition the effect of observed variables on mobility into direct and indirect (via satisfaction) routes but provide no information about how mobility could be altered by changing satisfaction. In terms of path analysis, satisfaction would be an endogenous intervening variable of little substantive impact. Only the reduced form equation relating mobility to objective variables would yield meaningful impact parameters.

The response to this objection is that satisfaction does depend on socially identifiable but missing or unobserved factors, which give it systematic exogenous variation. On the one hand, detailed case studies link job satisfaction to a host of very specific aspects of the work place, such as mode of supervision, physical work conditions, and so forth (see Locke; Victor Vroom) which are not generally measured on large data files, making satisfaction a potential proxy for those *unobserved objective factors*. On the other hand, lack of adequate information on the alternatives facing individuals makes the variable a reasonable indicator of alternative job opportunities, if those with good opportunities are less satisfied than those with poor opportunities. Some insight into the relative importance of omitted characteristics due to changes in the features of the current work place and of alternatives might be garnered from longitudinal information on changes in the job satisfaction and wages of mobile workers.

The omitted variable argument can be developed further by assuming that mobility depends solely on objective factors, including the omitted variables, and by treating satisfaction as an indicator of the omitted factors. If, as seems reasonable, the omitted aspects of the work place are

correlated with the measured factors, least squares estimates of their effect will be biased. Consistent estimates could be obtained by using satisfaction and other (subjective) variables that depend on the unobserved work characteristics as proxies, using general unobservable models. In this case, the satisfaction variable is needed to correct for econometric problems in estimating the effect of the observed variables. Whatever model structure is preferred, the link between satisfaction and objective but unmeasured variables rescues the satisfaction variable.

Finally, even if the interpretative problems with job satisfaction measures cannot be entirely resolved, the evidence that satisfaction is related to future mobility and other overt behavior (wages and standard variables held fixed) does provide useful clues to individual actions and to needed areas of research. It suggests that nonpecuniary factors are important in mobility and that additional effort be devoted to measuring and analyzing those factors.

IV. Job Satisfaction as a Dependent Variable

The definition of job satisfaction in industrial psychology as a "positive *emotional state* resulting from the appraisal of one's job" (Locke, p. 1300), highlights the principal problem in interpreting responses to satisfaction questions: that they depend not only on the objective circumstances in which an individual finds himself but also on his psychological state and thus on aspirations, willingness to voice discontent, the hypothetical alternatives to which the current job is compared, and so forth. Because job satisfaction reflects both objective and subjective factors, it is more complex than standard economic variables and requires more sophisticated and careful analysis. By altering the way in which persons respond to questions, variables like education (which raises aspirations) or collective bargaining (which provides a mechanism for "voicing" discontent) could have very different effects on job satisfaction than on objective economic conditions. The impact of satisfaction on overt behavior could also differ among groups, depending on the importance of objective and subjective factors in responses.

The distinct features of measured job satisfaction that result from its dependence on psychological as well as objective circumstances might be analyzed by comparing the effect of variables on satisfaction with their effect on overt mobility behavior (satisfaction excluded as an explanatory factor). Assuming that overt mobility depends *solely* on objective circumstances while satisfaction is influenced by subjective as well as objective factors, marked inconsistencies between the effect of variables on the two outcomes could be interpreted as reflecting the dependence of satisfaction on the subjective factor.

Estimates of the effect of various economic variables on job satisfaction (measured, as before, by a Z-score scaled so that positive values reflect increased satisfaction) and on the probability of quits (satisfaction held fixed) were made for the *PSID* and older male *NLS* samples. Because unionism was not available in the older male *NLS* until 1969, the calculations focus on quits from 1969–71. Table 3 summarizes the results in terms of the coefficients on variables having markedly different effects on satisfaction and quits.

The principal paradoxical finding is that trade unionism, which reduces quits significantly in the data sets, and thus would be expected to raise job satisfaction, either reduces it significantly (in the *PSID* and in the 1971 satisfaction equation in the older male *NLS*) or has little effect (1969 satisfaction in the older *NLS*). A negative or negligible coefficient of unionism on job satisfaction has also been found in other data sets (see James Hughes), including the younger male *NLS*, and has been documented, with a different model, for the older male *NLS* by Borjas. At the 1975 meetings, I suggested that the inverse relation might reflect the role of unions as a "voice" institution, encouraging workers to express discontent during contract negotiations and to make formal grievances

TABLE 3—ESTIMATES OF THE DIFFERENTIAL EFFECT OF UNIONISM AND JOB TENURE ON SATISFACTION AND QUITS

	Michigan *PSID*		*NLS* Older Male		
	Satisfaction 1972	*P*(Quit) 1972–73	Satisfaction 1969	Satisfaction 1971	*P*(Quit) 1969–71
Union	−.15	−.35	.04	−.13	−1.93
	(.04)	(.16)	(.05)	(.05)	(.42)
Tenure	−.001	−.06	+.000	−.002	−.16
	(.002)	(.01)	(.002)	(.002)	(.03)
R^2/ (*ln* likelihood)	.067	(2385)	.073	.075	(231)

Note: All equations include controls used in Table 2. The *P*(Quit) estimated on logistic function using maximum likelihood. Sample sizes, as in Table 2 except for older male *NLS*, which had 1,735 observations.

rather than to quit, which would keep the dissatisfied from leaving the employer. If this view is correct, the satisfaction relation lends some support to the exit-voice model of the union (see the author). Since wages are included in the calculation and since a negative relation is found for young as well as older workers, it is difficult to account for the anomolous relation in terms of the flatter age earnings profile of union workers, or related objective factors.

The other variable with consistently different effects is tenure, which is associated with much lower quit rates (possibly because of selectivity) but which has virtually no effect on job satisfaction. This could reflect the greater aspirations of those in a company due to increased benefits with seniority, their greater willingness to voice discontent due to job protection, or other subjective factors. While there were other differences in the effect of variables on satisfaction and quits in some of the data sets, there were no other clear patterns for all of the samples. Most variables like age, wages, and a race dummy had the expected opposite coefficients on satisfaction compared to quits.

Overall, the results of comparing satisfaction as a dependent variable with quits indicates that, consistent with economists' suspicion, satisfaction cannot be treated in the same way as standard economic variables. The divergent effects of unions (and to a lesser extent tenure) on satisfaction and quits suggests that at least some economic institutions and variables have very distinct effects on the subjective way in which individuals view their job satisfaction.

V. Conclusion

This paper has attempted to show that subjective variables like job satisfaction, which economists traditionally view with suspicion, contain useful information for predicting and understanding behavior, but that they also lead to complexities due to their dependency on psychological states. The empirical analysis has found job satisfaction to be a major determinant of labor market mobility and has turned up puzzling relations between certain economic variables, notably unionism, and satisfaction that appear attributable to the subjective nature of the variable.

REFERENCES

G. Borjas, "Job Satisfaction and Unionism: A Reappraisal of the Union Wage Effect," unpublished paper, Univ. Chicago 1977.

R. Flanagan, G. Straus, and L. Ulman, "Worker Discontent and Work Place Behavior," *Ind. Relat.*, May 1974, *13*, 101–23.

R. B. Freeman, "Individual Mobility and Union Voice in the Labor Market," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 361–68.

D. Hamermesh, "Economics for Job Satisfaction and Worker Alienation," in Orley Ashenfelter and Wallace Oates, eds., *Essays in Labor Market and Population Analysis*, Princeton 1977.

J. Hughes, "Satisfaction and Union Voice," undergraduate honors thesis, Harvard Univ. 1977.

E. A. Locke, "The Nature and Causes of Job Satisfaction," in Marvin Dunnette, ed., *Handbook of Industrial and Organizational Psychology*, Chicago 1976, 1297–350.

Victor Vroom, *Work and Motivation*, New York 1974.

"National Longitudinal Survey 1966–69," Ohio State Univ.

Survey Research Center, *A Panel Study of Income Dynamics: Study Design, Procedures, Available Data*, Ann Arbor 1972.

Psychic Income: Useful or Useless?

By LESTER C. THUROW*

Because work clearly provides opportunities for nonmonetary benefits and costs—fame, power, friends, physical discomfort, risk to life, etc.—economists traditionally add psychic income to money income and talk about total income maximization when analyzing labor supplies. In adding psychic income there is often the implicit assumption that psychic income makes little difference to the conclusions which follow or the analytical apparatus to be employed. As I shall show, neither of these implicit assumptions is warranted.

Before going on to analyze some of the impacts of psychic income, it is necessary to remember two fundamental propositions about psychic income. First, no job characteristics can be assigned a priori to the positive or negative component of psychic income. It all depends upon the relative supplies and demands for the characteristic. In a society where many people place a high premium on "machismo," a job with high physical risks may be considered a better job than a job with low physical risks. Similarly, in a "work ethic" society, working time may have a positive value while leisure time has a negative value. And whatever the aggregate market value of the characteristic, individuals may assign it a very different value. Second, a negative aggregate psychic income (the disutility of work) is not necessary for economic analysis. If one thinks of a conventional labor supply curve, the net utility or disutility of work simply determines the point where the supply curve intersects the quantity axis. If the job generates net utility, the supply of labor will be positive at a zero wage. If the job generates net disutility, the supply of labor will be negative at a zero wage and a positive wage will be necessary to produce a zero labor force. In either case monetary wages will be paid as long as the demand curve intersects the supply curve in the positive quadrant.

The standard technique for determining the value of psychic income is to estimate a wage equation that includes job characteristics (i.e., heat, etc.) as explanatory variables. The coefficients of these job characteristics are then used to provide an estimate of the amount of money (positive or negative) that would be equivalent to that characteristic.[1] Unfortunately, this technique is not apt to yield good estimates of the monetary value of different types of psychic income.

To illustrate the point, consider a case where a new positively evaluated job characteristic has been added to some job. In Figure 1 the job characteristic has a real value based upon the vertical shift (ab) in the labor supply curve, but the observed change in money wages ($w_1 - w_2$) is much smaller. Producer's welfare or psychic income is given by the rectangle $w_3 w_2 ab$ and the equivalent money wage is w_3. Basically, the more elastic the derived demand curve for labor, the less the observed wage change can be used to measure the psychic income received.

The estimation problem becomes even more complicated if we take into account the impact of providing psychic income upon the usual derived demand curve for labor. Psychic income can be generated in two ways. First, some job characteristics are simply an intrinsic condition of production. They would exist regardless of the preference functions of the work force. But second, some job characteristics (like the coffee break) are deliberately created to take advantage of the preference functions of potential workers. To provide this second set of job characteristics, the employer must incur extra monetary costs.

If the gross amount of money that can be given to labor is defined by profit maximiza-

*Sloan School of Management, Massachusetts Institute of Technology.

[1]For example, see R. E. B. Lucas.

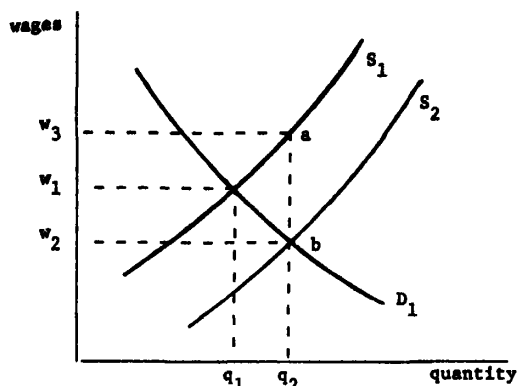



FIGURE 1

tion, then any decision to provide positive job characteristics or reduce negative job characteristics involves a leftward shift of the usual derived demand curve. Thus the observed change in wages is partly a function of the shift of the supply curve but partly a function of the shift in the derived demand curve for labor. In the case illustrated in Figure 2 a negative job characteristic has shifted the labor supply curve leftward from S_1 to S_2 and the employer's attempt to partly eliminate this undesired characteristic has shifted the derived demand curve for labor inward from D_1 to D_2. It should be noted that in this case there is a complicated interaction between the shifts in supply and demand. The amount that the supply curve will shift depends upon how

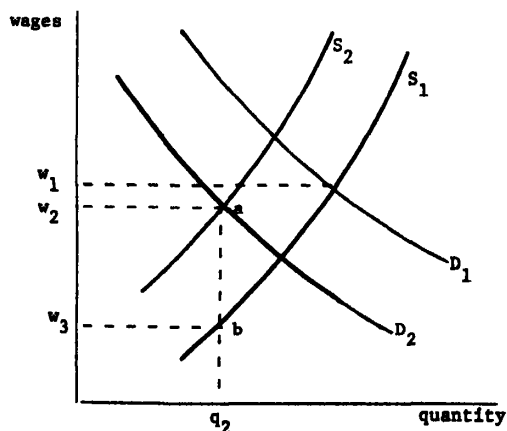



FIGURE 2

much the employer invests in eliminating the undesired characteristics, and the shift in the demand curve depends upon how much the supply curve would shift if the characteristic were not attenuated.

In this case, using the standard technique, the observed decrease in wages would lead one to think that the characteristics generated positive psychic income where it in fact generates negative psychic income. The individual earns a money income of $[(w_2)(q_2)]$, a negative psychic income of $[(w_2 - w_3)(q_2)]$, and a total income of $[(w_3)(q_2)]$. Thus, to calculate an employee's total income it is necessary to estimate both labor supply curves and derived demand curves with and without the characteristic.

The standard technique will also be biased for two other reasons. 1) If the marginal worker receives no psychic income, then psychic income has no impact on observed wages yet intramarginal workers may be receiving psychic income benefits and hence higher total incomes. 2) Suppose that the aggregate labor supply curve were perfectly inelastic and new positively evaluated characteristics were added to each job in the society. There would be no shift in supply curves, yet total income has increased.

The introduction of psychic income creates a host of problems for the rest of economic analysis. Many of the standard conclusions of economics depend upon the implicit assumption that individual utility and social welfare can only be created by consumption and not by production. Free markets and money income maximization can be shown to maximize the potential supply of consumption goods, but they do not necessarily maximize the aggregate supply of consumption goods plus producer's welfare. The standard economic calculus simply does not consider producer's welfare—area $(w_2 w_3 ab)$ in Figure 1 or Figure 2. (Note that producer's welfare is not a synonym for either consumer's or producer's surplus.)

Producer's welfare also undercuts the idea that workers should only be interested in wages and not worry about the tech-

niques of production chosen by his employer. If a shift in production techniques leads to a technique of production that yields fewer psychic benefits, the change should not be accepted unless the increase in wages can compensate, or more than compensate, for the lost psychic income. In Figure 1 the original work force would not voluntarily shift from technique 2 to technique 1 unless wages rose from w_2 to w_3, but market forces will only cause money wages to rise to w_1.

Another problem arises because the economy generates two currencies—psychic income and money income—where there are imperfect opportunities for trading the two currencies and where one of the two currencies only enters the utility function of workers and not that of employers or consumers. This is especially true in the case of the intrinsic job characteristics as opposed to the produced job characteristics.

Two imperfectly exchangeable currencies also create problems for the concept of equilibrium. An individual's net economic position cannot be reduced to one money number. Instead, his position has to be represented as a vector of benefits and costs with limited tradeoffs between the characteristics. As a result individuals may be out of equilibrium in the money area to be closer to their optimal position in the psychic area. Given that psychic income cannot be bought and sold as can money income, equilibrium becomes much more complicated both at a moment in time and over time. One type of psychic income cannot be exchanged for another form of psychic income and the time path of psychic income cannot be rearranged without rearranging production. Maximizing lifetime psychic income is very different than maximizing lifetime money income.

There are also enormous problems caused for the measurement of income distribution. Total earnings cannot be determined by multiplying the market wage rate by the quantity of labor employed, as seen in Figure 1. The welfare represented by (w_2, w_3, ab) simply lie outside of the conventional economic analysis. As a result, money income is not a good measure of the distribution of total income. The distribution of money income also depends upon the distribution of psychic income in addition to the distribution of gross marginal products.

Another major impact of psychic income occurs through the vehicle of interdependent preferences and the resultant importance of relative as opposed to absolute income in making work decisions. Substituting interdependent preferences for independent preferences would not make much difference to the theory of labor supply by itself (employers would simply ignore the interdependent preferences just as they ignore the efficiency with which individuals generate utility from income in a world of independent preferences), but it makes a great deal of difference if some of the other counterfactual assumptions of classical labor supply are lifted at the same time.

Basically there are three other major counterfactual assumptions: skills are acquired exogenously in formal education and training and then sold in an auction labor market; individual workers have fixed marginal products which do not depend upon their psychic evaluation of their circumstances; and in a world without economies or diseconomies of scale, total production is simply the sum of individual productivities.

While this is not the place to argue the validity of these assumptions at length, there is substantial evidence that all four of the standard assumptions are counterfactual. Surveys of work satisfaction, the importance of wage contours, and the sociology of relative deprivation all point to the existence of interdependent preferences.[2] Surveys of where working skills are acquired show that they are mostly acquired endogenously on the job in an informal transmission process from one worker to another.[3] The entire thrust of industrial psychology revolves around the question of getting workers to provide their highest marginal product rather than some lesser amount.[4] Finally, there is the prob-

[2]See Lee Rainwater, Richard Esterlin, Walter Runcimen, and John Dunlop.

[3]See U.S. Department of Labor.

[4]See Edward Lawler.

lem of team as opposed to individual productivity. The importance of team productivity and its interaction with relative wages has been a leading topic on most sports pages, but it is of no less importance in other industrial operations. Learning curves can be seen as an expression of how productivity dramatically rises as teams of workers polish their individual and collective skills.[5]

If skills are acquired endogenously on the job, if individuals can vary their productivity depending upon job satisfaction, and if team work is important, interdependent preferences become operational in the sense that employers must take them into account in setting wages. If these interdependent preferences are not satisfied, workers can stop training, lower their own productivity, and disrupt the team. In theory, such actions could be prevented by an all-knowing employer who would reduce wages and/or fire the appropriate worker. But this ignores the costs of hiring and firing, the costs of finding out who has or has not reduced their productivity, and the fact that such firings themselves would disrupt teamwork. The need for voluntary cooperation is an area which economics ignores. Yet in the end all production depends upon such cooperation. Working to rule can bring any operation to a halt.

Team productivity also leads to a payment—the excess of team productivity over the sum of individual productivities—that must be awarded on the basis of some principle that is not determined in classical marginal productivity theory. Presumably to keep the team working in harmony, these extra payments must be distributed in accordance with the team's norms of industrial justice. The importance of interdependent preferences is also enhanced by endogenous on-the-job training since wages cannot be determined in the Marshallian supply and demand scissors. If training is only given when a job opening exists, then the supply curve of labor is a function of the demand curve for labor. At the extreme, the supply curve for labor lies on the demand curve for labor and any wage rate is an equilibrium wage rate. As a result, the interdependent preferences of the labor force do not in general have to bump up against alternative market wages.

From this perspective psychic income can help explain many of the observed realities of the labor market. Rigid relative wages and the downward rigidity of money wages can be seen as an expression of interdependent preferences and the psychic income which they generate. In this context rigid wages actually exist, they may be a necessary condition for dynamic efficiency. Without them productivity would fall and training would come to a halt. Similarly, the team payment can provide a good explanation of why it seems impossible to find the equilibrium skills wages which simple economic theory uses. In addition to their personal characteristics, individuals are paid based on whether they play for a high productivity or a low productivity team.

[5]See Steve Shreffrin and the author.

REFERENCES

John Dunlop, *Wage Determination Under Trade Unions*, New York 1950.

R. Esterlin, "Does Money Buy Happiness?," *Publ. Interest*, Winter 1973, No. 30, 3–10.

Edward E. Lawler III, *Pay and Organization Effectiveness: A Psychological View*, New York 1971.

R. E. B. Lucas, "Hedonic Wage Equations and Psychic Wages in the Returns to Schooling," *Amer. Econ. Rev.*, Sept. 1977, *67*, 549–58.

L. Rainwater, "Poverty, Living Standards and Family Well-Being," work. paper no. 10, Harvard-Mass. Instit. Technology Joint Center Urban Stud. 1974.

Walter G. Runcimen, *Relative Deprivation and Social Justice*, London 1966.

S. Sheffrin and L. C. Thurow, "Estimating the Costs and Benefits of On-the-Job Training," *Econ. Appliq.*, forthcoming.

U.S. Department of Labor, *Formal Occupation Training of Adult Workers*, Manpower Automation Research Monograph no. 2, Washington 1964, pp. 3, 18, 20, 43.

DISCUSSION

RUDY OSWALD, AFL-CIO: The basic theme that runs through all three papers is that job satisfaction is a necessary consideration for economic analysis. But what is job satisfaction?

Is it some set of responses to satisfaction questions, or is it "quit rates" (Richard Freeman); is it employee participation in decision making, or "social learning" (Karl-Olof Faxén); or is it "fame, power, friends, physical discomfort, risk of life, etc." that Lester Thurow calls psychic income?

Maybe it is all of these things, plus much more—more than can easily be measured with existing tools. In general, job satisfaction is approached by the behavioral scientist as either a psychological study of workers classified into such standard variables as age, sex, race, income, and related factors, or as an organizational behavior tool to be used to increase productivity. I maintain that neither of these approaches identify a worker's satisfaction on the job. They disregard the basic needs of workers for income and security—and the role of unions in our society in trying to assure these basic needs. Indeed I find it shocking that these representatives of the economic profession can describe various aspects of job satisfaction without taking account of the role of unions in our economic system.

Pay itself is directly related to job satisfaction. Only Faxén emphasized the "overwhelming importance of pay," while Thurow recognizes the need for "positive wages." These analyses assume that the "pay" for the job meets some level of sufficiency, and disregard pay above that level as a measure of satisfaction. Yet for many workers their pay has not yet attained a level of sufficiency that is a prerequisite for consideration of that luxury called "job satisfaction." Secondly, only Freeman mentions the importance of job security as a criteria of job satisfaction. Without job security, or the ready prospect of alternative job opportunities, job satisfaction is again a luxury.

Unions, representing the concerns of their members, place primary emphasis on these two basic elements of job satisfaction—adequate wages and job security. First in regard to wages, through the collective bargaining process, wages are no longer solely determined by the employer, but the result of a codetermination of the appropriate wage. This is an important element that is neglected in these papers' concentration on aspects of satisfaction not related to wages or the importance of workers having an equal voice with employers in establishing wages. Further studies of job satisfaction should analyze both the level of wages and the role of the worker in determining that level.

Second, job security has a direct relationship to quit rates and turnover as Freeman points out. In this paper he only alludes to the importance of the various types of job security that are provided under most union contracts. The grievance procedure provides an outlet for resolving problems, the discharge and discipline clauses assure certain intrinsic elements of justice and fair play, the seniority clauses assure protection against favoritism, the 540 clauses and attrition clauses protect incomes, and many of the other parts of contracts provide job security. These written contract clauses are clearly related to job satisfaction.

Let me hasten to add that unions frequently seek the broad type of job satisfaction discussed in these papers. Many strikes are generated by job dissatisfactions. One example that comes to mind is the legend written on the picket sign of one of the garbage collectors in that famous Memphis strike of 1968. That striker wrote the simple legend "I am a Man." And in many of the issues involving work that is still an issue today. Faxén talks about "worker participation" but he ignores the meaning of "unionism" and the role of the union as an organization of workers in "participation." While Faxén adds to the documentation of the substantial contribution made by "job satisfaction" to

increased labor productivity, and Thurow's paper provides a theoretical basis for such results, *U.S.* employers seem to be interested in job satisfaction more as a device to circumvent or destroy income.

Today, *U.S.* employers seem to be engaged in substantially greater anti-union activities than exhibited in many a decade, while paying lip service to the notions of increasing worker's job satisfaction and labor productivity. One of the glaring examples is the National Association of Manufacturers drive for a "union free" environment that would inhibit workers from having a voice in wage setting, job security, and other aspects of job satisfaction.

Is this new employer attitude an appropriate environment to improve the quality of working life and enhance job satisfaction? Or are we entering a new era, where job satisfaction in 1984 terminology is just the opposite? I maintain that a harmony needs to exist between workers represented in collective bargaining and employers for there to be true job satisfaction. Otherwise job satisfaction means worker manipulation.

GEORGE STRAUSS, University of California-Berkeley: As one whose major identification is organizational behavior, I rejoice that at last the American Economic Association has a session dealing with the quality of work life. I rejoice not because of sinners saved, but also because, compared to the somewhat murky discussions among psychologists, economists have an ability to state propositions with succinctness, elegance, and exactness; characteristics particularly evident in the papers here.

Karl-Olof Faxén's paper provides useful data as to Swedish developments. I am sorry only that space did not permit more detailed description of the research sites and research methodologies. His findings contrast with those of American studies which frequently show quality of work life experiments leading to higher morale and easier introduction of change, but rarely demonstrate increases in productivity as such. Of course, quality of work life changes often require "social learning" to make them work; frequently they also require investment in training and new equipment. Thus it may be oversimple to call them purely "disembodied."

There have been literally thousands of psychological studies relating satisfaction to turnover. Nevertheless Richard Freeman's paper distinguishes itself by its sophistication. Ironically it comes at a time when the concept of job satisfaction is under massive attack by psychologists on both methodological and conceptual grounds. They argue, for example, that 1) single question measures of satisfaction (such as those utilized by Freeman) are inadequate and instead attitudes toward specific job facets such as pay, boss, fellow workers, and nature of task should be measured separately; 2) job satisfaction data alone are insufficient unless *importance* of job is also measured; 3) reported job satisfaction depends on numerous factors other than the job itself, for example, how the question is asked, attitudes on one's fellow workers, one's aspriations, one's perceived opportunities for change, etc.; 4) pay itself is not independent of job satisfaction, satisfaction with so-called *extrinsic* factors (such as pay) is dependent in part on satisfaction with intrinsic *factors* (such as job challenge); for example, increased pay, *ceteris paribus*, may lead to increased, decreased, or no change in satisfaction with job challenge, depending on the contextual circumstances. At the least economists working on this area should immerse themselves in the satisfaction literature. Further, since the basic variable is somewhat ill-defined, conclusions probably should not be drawn from small differences in satisfaction.

Lester Thurow's paper opens new horizons. Economic theory can help understand not only increased worker's preference for nonpecuniary (for example, agreeableness of work) as opposed to pecuniary goods (for example, pay) but also attempts by companies to redesign jobs to

adjust to these changed preferences. For a given total cost per unit of output, jobs can be designed to provide workers various combinations of pecuniary and non-pecuniary rewards. An iso-cost curve can be constructed to show this, and geometric analysis (too complex to present here) suggests that the optimum degree of job enrichment lies somewhere between the point where further enrichment will reduce productivity and the point where it will reduce satisfaction. Thus productivity should not be the sole criterion for evaluating quality of work life efforts.

The Costs of Inflation

By GARDNER ACKLEY*

It is a famous proposition in economics that fully anticipated inflation (or deflation) has no significant effect on the level or distribution of real income. As with other such facile generalizations, this proposition seems both true and essentially meaningless. Clearly, such a benign inflation must have remained at a steady rate over a considerable period in order to be fully anticipated. In fact, it must have been anticipated long before it began, at least by parties to long-term contracts such as fifty-year bonds. However, there has never been a steady rate of inflation for any appreciable period in any country, and evidence has never been presented that *any* inflation was accurately anticipated—even on balance—while there is much (admittedly casual) evidence that particular inflations have been unexpected by great numbers of those affected.

There is clear evidence that the variability of inflation increases with its average extent. Arthur Okun (1971) showed that the higher was a country's inflation during 1951–68, the more variable it was, and thus the less likely to have been anticipated; Dennis Logue and Thomas Willet reached similar conclusions using data on a larger number on countries and over a longer period. Benjamin Klein showed that the variability of *U.S.* price change, 1880–1972, was highly correlated with the moving average of the absolute annual rate of price change (ignoring sign). Only in the years 1955–72 was this correlation significantly disturbed—by an unusually low variability of the inflation rate. (An extension of Klein's calculations for the years 1973–77 shows the variability of the inflation rate somewhat increased, but still quite low by earlier standards and relative to the average rate; but changes in its variability were still correlated with changes in the average rate.)

In my view, the principal significance of the standard proposition is not the usual conclusion that steady (and thus presumably expected) inflation is costless, but rather its implication regarding the *special and additional* costs of the initiation or acceleration of inflation. Doubtless, there are also special costs of an unexpected reduction in the rate of inflation; but I will argue that any continuing positive average rate of inflation involves costs that are the more serious the higher the rate. Thus there is a clear net gain from reducing the average rate of inflation, provided that the costs of securing that reduction are not too great and that the lower rate is maintained.

I. The Redistributive Effects of Inflation

The costs of inflation are of two broad kinds: redistributions of income and wealth that serve no economic purpose; and reductions in the level or rate of growth of production. These costs are more significant the more unexpected the inflation, but they occur even if inflation is largely expected. Redistributions of income and wealth, stemming from changes in relative prices, of course arise even under a stable price level, when they are appropriately regarded as a necessary means for adjusting the structure and mix of productive activities to changing technology, new products, altered tastes, and so on. Such redistributions are multiplied when the general price level changes, for two broad classes of reasons: first, because there are uneven lags in the response of individual prices and wages to the forces that permit or create inflation; and second, because, unavoidably, many important economic arrangements must extend over time.

*Professor of economics, University of Michigan.

The differential lags in the response of particular wages and prices are mostly the result of the fact that in modern economies almost all wages and the bulk of all prices are "administered" rather than market clearing. They change only periodically, through discrete decisions made by individuals or groups that possess varying degrees of market power (and hold particular price level expectations). Those who decide are applying their particular "policies"—practical rules of thumb, usually reflecting some concept of "equity." Those responses are never instantaneous, and there is no reason for them to occur at a uniform speed. It is likely that, under the stress of prolonged and substantial inflation, these institutional response patterns evolve in the direction of compressing all lags toward zero. In my view, this mainly causes the rate of inflation whatever its source to accelerate, probably without much reduction in the redistributive effects and perhaps even increasing them.

Indeed, some of us believe that an important force making our economy so inflation prone is the ability of many organized groups in our society to obtain—through market or political action—changes in relative prices that are expected to be favorable to them, although these expectations are repeatedly frustrated through the delayed rise in other prices. Thus, changes in relative prices are an integral part of the means by which inflation is propagated.

The other main reason why inflation creates changes in relative prices that redistribute income and wealth is the inevitability of economic arrangements that extend over time. Everyone who enters into contracts calling for future payments or receipts makes a bet on the future that requires some kind of price level expectation, consciously formulated or not. Indeed, such bets are implied by almost every future-oriented economic decision—for example, to change residence, to prepare for a career, to buy a durable physical asset—often where no explicit contract is involved, or is even possible. When these expectations fail to be fully realized, unexpected (and therefore "unjust") changes in the distribution of income and wealth necessarily occur.

The high productivity and resulting high incomes of modern economic society require that people make many such bets. They must participate in large impersonal markets, and make actual or implied contracts for the continuing sale of their services or products; they must produce and own highly durable instruments; they must participate in capital markets that separate the acts of saving and investment; individuals, firms, and governments must *plan* their activities and thus make or require commitments (or impose obligations) which continue *over time*.

The redistributions of income and wealth which arise from these arrangements in an inflationary world are most serious when formal or implicit commitments cover the longest periods of time. Yet much of our productivity stems from the use of highly durable goods, which require extremely long-term financing (Con Ed is not going to finance a nuclear generating plant with commercial paper and bank loans—nor with equities), and often long-term contracts for sale of outputs or purchases of inputs. Indexation of long-term debts and contracts may be part of a "second best" answer to inflation. However, even with indexation a world of inflation involves greater uncertainties for lenders and borrowers—and thus greater real costs of production and lower real incomes—than does a world without inflation.

It is important to know whether the redistributions associated with inflation systematically involve broad social or economic classes or groups (suppliers of labor vs. suppliers of property services, organized vs. unorganized workers, farm vs. urban communities, and so on). G. Leland Bach, Bach and Albert Ando, and Bach and J. B. Stephenson provide evidence that few such systematic effects do appear. The principal broad class of losers, both on income and wealth account, seems to consist of retired persons whose wealth and income are primarily associated with direct

or intermediated ownership of fixed-income securities. However the relationship between the extent of inflation and of changes in relative prices involving narrower economic groups should be the object of more research. Daniel Vining and Thomas Elwertowski recently showed a strong positive correlation between the general inflation rate and the dispersion of inflation rates for particular goods and services, suggesting that inflation may imply greater income redistributions of this type than those that occur with price-level stability.[1]

All income redistributions, whether among classes or individuals, increase personal insecurity and lessen personal satisfactions (even on the part of the beneficiaries), and heighten interpersonal and institutional tensions. Thus they are destructive of the social and political fabric, and ultimately of economic efficiency. Even those who are not really hurt by inflation often think they are. They regard their gains in money incomes as the result of their own cleverness, the effectiveness of their union or trade association or friends in political power—or just good luck. These benefits would have occurred whatever happened. The rise in the prices they pay therefore is regarded as unjustly robbing them of what is rightfully theirs. Thus they think they are hurt; and if they think they are, they are. All seek someone to blame—those greedy employers or nasty trade unions; the bankers, landlords, farmers, or foreigners; our economic system, the government, society. A significant real cost of inflation is thus what it does to morale, to social coherence, and to people's attitudes toward each other.

[1]Unfortunately, the Vining-Elwertowski evidence is not conclusive. The author's own analysis of relative price changes during the Korean War inflation showed that the very sharp increase in the dispersion of wholesale price changes at that time (and the subsidence of the dispersion as the inflation rate declined) was largely associated with the degree of processing. The fact that prices of hides might rise 100 percent, of leather 80 percent, of shoes at wholesale 20 percent and at retail 15 percent does not necessarily reflect major income redistributions among participants in the process (other than to hide producers).

II. Effects of Accounting and Taxes

The effects of inflation on the distribution of income and wealth are of course exaggerated by conventional methods of business accounting. Managers, stockholders (and potential purchasers of stocks), lenders, public officials, and others could in principle all make their own private inflation adjustments of the public accounting records of firms in which they are interested, in order not to let their decisions be warped. In fact, most of them have no idea how to do this, and (except perhaps for the managers) lack most of the detailed data necessary to do it. Indeed, the purpose of public accounting is precisely to supply useful information to all of these groups, and the destruction of such information is one of the real costs of inflation.

As Henry Aaron's study documents, the situation is made worse by the fact that these same private accounting records are used by the government for levying business taxes (and sometimes for price controls, especially of utility rates). Most of the inadequacies of current accounting lead to a serious increase in the real taxation of business income during inflation. Moreover, the extent of these tax increases is quite uneven among firms and industries, depending on their particular asset and liability structures, methods of inventory valuation, and so on. Quite unintentionally and arbitrarily, in the presence of inflation, taxation based on standard accounting not only redistributes income, but may well reduce and distort investment incentives. For example, Martin Feldstein recently showed that, in the presence of inflation, taxation of interest income can distort the saving rate and capital intensity of a Tobin growth model.

The effects of inflation on the taxation of personal income are similar to those on corporate income, insofar as income from unincorporated enterprises is concerned. The existence of fixed deductions, exemptions, and tax brackets in personal income taxes means that inflation also raises real rates of taxation on individuals—unintentionally and arbitrarily—thus tending to

reduce real incomes and to depress aggregate demand. To the extent that government budgets are fixed in money terms, with inadequate allowances for inflation, a rising price level also reduces real government purchases and, again, aggregate demand.

It would be nice if there were some single, new, all-purpose set of accounting conventions that would resolve these problems. There are proposals that would reduce particular distortions; but each leaves other distortions untouched, and often creates new ones. Again, accounting reform is part of a second best solution—which may reduce but will not eliminate some of the redistributional costs of inflation. The same is true of tax indexation.

III. Effects on Aggregate Demand

Any textbook in macroeconomics accumulates a considerable list of probable effects of inflation on *aggregate demand*, and thus, *ceteris paribus*, on total production and incomes. Here are seven familiar ones, the first five of which are negative, and occur whether or not the inflation was expected. The last two are positive and relate mainly to expected inflation.

By increasing the nominal demand for money, inflation forces up interest rates, deterring investment.

By reducing the real value of aggregate consumer wealth (to the extent composed of government debt and money), it inhibits consumer spending.

By raising effective tax rates and reducing real government purchases, it makes fiscal policy more restrictive.

By raising domestic prices relative to foreign, it inhibits exports and stimulates imports.

By increasing consumer fear and feelings of insecurity, it increases the propensity to save.

By increasing expected future prices of output relative to current costs of capital goods, it encourages investment.

By encouraging purchases now instead of [illegible] it moves investment and consump[illegible]ward in time.

If an inflation is caused by excessive aggregate demand, or—however caused—if it can be effectively reduced through curtailing aggregate demand, inflation's negative effects on demand may be welcome, despite their cost in employment and output. But if the inflation was not caused by excessive demand, or is not significantly controllable through reduction of demand, its effects on output and incomes are unwelcome.

IV. Effects on Aggregate Supply

The more interesting questions regarding the effect of inflation on aggregate output, however, relate not to demand effects but to impacts on supply, which are not easily offset, and are never welcome. Essentially, they are effects which reduce the total output (or the value to consumers of the mix of output) that can be produced with any given input of productive resources. They include the following:

1) Inflation, or at least its expectation, tends artificially to increase the production of goods as opposed to services, of more durable as opposed to less durable goods, and often of physical as opposed to human capital. This occurs because these goods are seen to cost less today than they will tomorrow and they will still be there tomorrow. In a sense, this is the same as the last demand effect listed; but it is seen here in its aspect as a distortion of production, and thus as a reduced real value of production to consumers. More concretely, firms are encouraged to carry larger inventories than they really need and to build plants and buy equipment sooner than really necessary. In countries with continuous rapid inflation, incomplete buildings are visible everywhere, awaiting gradual completion as their owners can gradually finance it; new machinery rusts awaiting use, coal piles are far larger than needed. These are gross examples of inefficiencies created by inflation; more subtle ones must be everywhere.

The economist may say that this behavior reflects the absence of suitable financial investments that could produce the same yield—perhaps because of ceilings on interest rates, the neglect of indexa-

tion, or the absence of institutional structures for providing alternate forms for the real investment of savings in units small and secure enough for low or moderate income savers. However these are not easy—and are often not economical—to provide.

2) Money and other deposit balances are excessively economized, requiring more frequent settlement of accounts than is reasonable, and so on, in order not to hold depreciating money. Again, this particular problem may be solved by removing interest ceilings on demand and savings deposits. But in some economies, demand deposits are not used by most of the population; and it is quite impossible to pay interest on holdings of cash.

3) Scarce managerial talent is diverted from managing production, maintaining efficiency, seeking economy, and innovating, in favor of maneuver, speculation, and the search for protection against (or benefit from) inflation.

4) Because long-term contracts of all kinds involve more risk in inflationary periods and places, people refuse to enter upon them, sacrificing the many real production efficiencies and economies which are made possible by such contracts, as well as wasting resources in more frequent negotiation.

5) Inflation destroys or weakens the usefulness of all kinds of market information which people accumulate merely through repeated transactions at a stable price level. With inflation, every transaction requires new information gathering, again to find the cheapest source, the most suitable quality at the price, and so on. Shopping for the family or the firm is made more difficult, as when one suddenly begins using a foreign currency. What is a "good buy" requires laborious calculations of what the price means in the familiar currency, and how it compares with prices of other goods—which requires finding out what other goods are selling for before deciding to buy the first—as Okun (1975) so effectively pointed out.

Economists do not know how to measure the full extent and significance of these and other effects of inflation on distribution and production. Clearly, they occur; yet they do not necessarily stand in the way of reasonable economic performance or even of economic progress. Some countries have lived with high and fluctuating inflation rates of 30 to 150 percent or more for decades, without complete breakdown or even highly visible impairment of production and living standards; some have even been able to achieve economic growth despite high inflation. I, however, would hold that the immensely more complex and completely market-oriented economies of the major Western countries, with their greater dependence on long-term contracts, and their vastly more intricate financial markets and intermediation, could not survive as well under these conditions.

Inflation tremendously complicates the task for the makers of government fiscal and monetary policy. Even if they believe that the costs of moderate and not too unsteady inflation are vastly overrated, the public does not. Thus, inflation not only makes it harder for policymakers to diagnose the factors affecting aggregate demand, but it also forces them to do—and even more often to *say*—silly things: for example, that inflation is intolerable, but in the next breath, that each of the things that might be considered to deal with it (other than to allow the inflation to create unemployment) is unthinkable because they involve some *cost*—as though inflation and unemployment did not! Yet, at the same time, policymakers can take other actions that clearly raise prices, a fact which they cannot afford to mention. Thus one of the worst evils of inflation is the accompanying deterioration of the level of public discourse.

Finally, now that almost all economists of every school at last agree that the main cause of inflation is past inflation, clearly the greatest cost of inflation is the inflation it causes.

REFERENCES

Henry Aaron, *Inflation and the Income Tax*, Washington 1976.

G. Ackley, "Selected Problems of Price Control Strategy, 1950–1952," National

Archives Microfilm, T460 Row 1, 1953.

G. L. Bach, "Inflation: Who Gains and Who Loses," *Challenge*, July/Aug. 1974, *17*, 48-55.

——— **and A. Ando,** "The Redistributional Effects of Inflation," *Rev. Econ. Statist.*, Feb. 1957, *39*, 1-13.

——— **and J. B. Stephenson,** "Inflation and the Redistribution of Wealth," *Rev. Econ. Statist.*, Feb. 1974, *56*, 1-13

M. Feldstein, "Inflation, Income Taxes, and the Rate of Interest: A Theoretical Analysis," *Amer. Econ. Rev.*, Dec. 1976, *66*, 809-20.

B. Klein, "The Social Costs of the Recent Inflation: The Mirage of Steady 'Anticipated' Inflation," *J. Monet. Econ.*, suppl. series, 1976, *3*, 185-212.

D. E. Logue and T. D. Willett, "A Note on the Relation between the Rate and Variability of Inflation," *Economica*, May 1976, *46*, 151-58.

A. M. Okun, "Inflation: Its Mechanics and Welfare Cost," *Brookings Papers*, Washington 1975, *2*, 351-90.

———, "The Mirage of Steady Inflation," *Brookings Papers*, Washington 1971, *3*, 485-98.

D. R. Vining, Jr. and T. C. Elwertowski, "The Relationship between Relative Prices and the General Price Level," *Amer. Econ. Rev.*, Sept. 1976, *66*, 699-708.

The Private and Social Costs of Unemployment

By MARTIN FELDSTEIN*

We do not need a careful quantitative analysis to establish that the unemployment of seven million workers involves very substantial private and social costs. Why then should we bother to think about measuring the cost of unemployment? There are two quite different reasons. First, by measuring the private costs of unemployment that are borne by the unemployed themselves, we can better understand why our unemployment rate is so high. Second, by examining the social costs of unemployment (i.e., the costs of unemployment to the nation as a whole regardless of how they are distributed), we can better decide when the benefits of a reduction in unemployment outweigh the costs of achieving it. The present paper considers both of these problems, emphasizing the conceptual issues rather than presenting specific estimates.

Since unemployment is so often thought of in aggregate terms, it is worth emphasizing at the outset that a proper analysis of the costs of unemployment must begin by disaggregation. The private cost of unemployment is very large for some of the unemployed, but is quite small for many others. The average private cost of unemployment is therefore much less relevant than the distribution of such costs. Similarly, in considering the social costs of unemployment, it is important to distinguish several kinds of unemployment since the cost of each type of unemployment and the costs of reducing that unemployment differ significantly.

I. Private Costs

The cost of unemployment that is borne by the unemployed person himself varies from the overwhelming to the trivial. At one extreme is the very substantial loss by those who experience a long period of unemployment with little or no help from transfer payments. At the other extreme is the minimal loss of those who are out of work very briefly and whose lost net income is fully replaced by unemployment compensation. Although there is a wide range of experience, the typical spell of unemployment is closer to the low cost extreme than to the high cost extreme. Even now, more than half of the unemployment spells last four weeks or less. Moreover, more than half of the unemployed received unemployment compensation. I believe that the relatively low cost of unemployment in these circumstances is a substantial cause of our high permanent rate of unemployment.

The principal reason for the low private cost of unemployment is the interaction of our tax and unemployment compensation systems. It is particularly important to consider these two together. Income and social security taxes now imply a marginal tax rate in the neighborhood of 30 percent for a worker in a median income family. It is therefore very significant that unemployment compensation is not subject to tax. The combination of a relatively high marginal tax on earnings and no tax on unemployment compensation implies that unemployment benefits replace a very high fraction of lost net income, typically about two-thirds.

An example will illustrate how this occurs. Consider a worker in Massachusetts in 1977 with a wife and two children. His gross earnings are $140 per week while hers are $100 per week. If he is unemployed for ten weeks, he loses $1400 in gross earnings but only $279 in net income. Why does this occur? A fall in gross earnings of $1400 reduces his federal income tax by $226, his social security tax by $82, and his Massachusetts income tax by $75. Thus, total taxes fall by $383, implying that net wages are reduced by $1017.

*Harvard University and the National Bureau of Economic Research. I am grateful to the National Science Foundation and the NBER for support of my research. This paper has not been reviewed by the Board of Directors of the NBER.

Unemployment benefits are 50 percent of his wage plus a dependents' allowance of $6 per child per week. The benefit is thus $82 a week. Since there is an annual one week "waiting period" before benefits begin, nine weeks of benefits are paid for the ten week unemployment spell. Total benefits are thus $738. The loss in net income is only the $279 difference between these benefits and the fall in after-tax wages. The $279 private net income loss is less than 20 percent of the loss in output as measured by the gross wage.

Because of the one week waiting period, the private cost of unemployment is even lower for an additional week of unemployment. If he stays unemployed for eleven weeks instead of ten, he loses an additional $140 in gross earnings but only $16 in net income. The private net income loss is less than 12 percent of the loss in output as measured by the gross wage. If the individual values his leisure and nonmarket work activities at even 50 cents an hour, there is no net private cost of unemployment!

The great reduction in the private cost of unemployment that results from this interaction of high taxes on earnings and the untaxed unemployment benefits produce substantial adverse incentives that magnify the cyclical volatility of unemployment and raise the noncyclical "baseline level" of unemployment. The most obvious effect is to increase the average duration of unemployment spells. With little or no personal cost of a longer period of unemployment, it is rational for the individual to look for a new job until the potential gain from additional search is extremely small or to use the low cost time to do chores at home or just to enjoy a period of vacation. In addition to increasing the average duration of existing unemployment spells, the low private cost of unemployment also causes an increase in the number of unemployment spells. Since workers who quit their jobs are eligible for benefits in a number of states, the low private cost of unemployment is responsible for many of the one million unemployed who quit their last job.

More significant, however, is the incentive for temporary layoffs. Approximately half of the unemployment spells that are officially classified as "job losses" are actually temporary layoffs in which the unemployed worker expects to return to his original job. In manufacturing, approximately 80 percent of those who are laid off return to their original jobs. Our system of unemployment compensation lowers the cost of such temporary layoffs to both firms and workers, making unemployment more attractive than accumulating inventories or cutting prices.

I have concentrated these comments on unemployment compensation because this is the most significant program for reducing the private costs of unemployment. Those who are not eligible for unemployment compensation often receive other forms of income replacement such as food stamps, social security, and welfare. It is also important to remember that a very large fraction of the unemployed who do not receive unemployment compensation are young people who are supported by their families.

It is easy to see how our system of taxes and transfers drastically lowers the relative private cost of unemployment and thereby induces a higher unemployment rate. The real puzzle is why the low private cost of unemployment does not result in a higher rate of unemployment. What are the forces of self-restraint that limit the public's willingness to exploit the full opportunities for subsidized unemployment? And will they continue to be effective in the future? Public attitudes about accepting transfer payments appear to have been changing rapidly during the past decade, resulting in the rapid growth of such things as disability insurance benefits, food stamps, and health insurance payments. The contagiousness of social attitudes suggests that this trend may accelerate in the future. It carries with it an ominous prospect for unemployment.

II. Social Costs

The social cost of an unemployment spell depends on the social opportunity cost of the unemployed person's time. In measuring the social cost of unemployment, it is

therefore crucial to ask "Unemployment as compared to what?" As economists, we tend to define the opportunity cost of any resource as its value in the best possible use to which it might be put. But the relevant opportunity cost in the current context is not this "best allocation" of full employment general equilibrium theory. We are interested in the social costs of unemployment in order to assess the desirability of particular unemployment policies. Different policies imply different opportunity costs for the unemployed workers. In each case, we should compare the particular net social cost of unemployment—that is, the potential benefit of returning the unemployed person to work—with the cost of the policy itself.

The format of this session suggests that all policies to reduce unemployment entail increasing inflation, implying that the relevant comparison is between the social costs of unemployment and the social costs of inflation. If this were true, the implication would be quite dismal since most economists now agree that a permanent increase in inflation cannot achieve more than a temporary reduction in unemployment.[1] Fortunately, there are policies for reducing unemployment permanently that do not involve increases in inflation. These policies may involve such costs as a reallocation of some workers from more productive to less productive activities, a reduction in unemployment insurance protection, or a redistribution of income with losses by some groups and gains by others. A proper evaluation of available policies requires quantifying the costs and benefits of each.

Before looking at some examples of the social costs associated with different types of unemployment, it is useful to comment on two extreme but common views of the social cost of unemployment. According to one view, unemployment has no social or private cost. The individual's loss of wage income is at least offset by the value of his leisure and of the information that he acquires by his job search activity. This conclusion is false even if we accept its premise that all unemployment is voluntary. The taxes and unemployment insurance described above imply a substantial gap between the individual's gross wage and the value of his time when unemployed. The existence of the rigidities that cause involuntary unemployment only strengthens the reason to reject this view.

At the other extreme is the view that the loss in wage income is equal to the social cost of unemployment. This ignores the value of the individual's leisure and of the information gained by searching for a new job. Moreover, even if both of these were zero, it would be wrong to regard the individual's normal or potential wage as a measure of the gain that would result from his reemployment without specifying the policy that would be used to achieve his reemployment.

Consider, for example, the case of workers on temporary layoff. As I noted above, some 80 percent of workers who are laid off by manufacturing firms soon return to their original jobs at those firms. Such temporary layoffs are almost completely unknown in Europe and Japan. This important source of unemployment could be significantly reduced if the employer tax that is used to finance unemployment compensation were changed to eliminate the current subsidy of excessive layoffs by some firms. While I believe that this would be a worthwhile reform, the benefit of such a change should not be overstated. The social cost of the unemployment that would thereby be eliminated is not the normal wage of these workers or even that wage reduced by the value of their leisure. A reduction in temporary layoffs would mean more production for inventory and more spells of below average productivity.

This example also illustrates the familiar principle of welfare economics that it may be possible to identify a good policy in terms of the marginal conditions without explicitly evaluating the gains from the policy. In this case, it seems clear that

[1] In a growing economy, the present value of the social cost of a permanent increase in inflation can be extremely large relative to the gain from a temporary reduction in unemployment; see the author.

eliminating the subsidy that increases temporary layoff unemployment would be a move in the right direction even though the value of the gain is unknown. Although the theory of the second best cautions against this general line of reasoning, an explicit partial equilibrium calculation of the gain from reducing unemployment is unlikely to be an improvement in this regard. An explicit calculation of the social cost of temporary layoff unemployment would be of value primarily in deciding whether the economic gains of the reform outweighed the political costs of achieving it.

Although the potential wage will generally overstate the social cost of unemployment, there is an important case in which it is an understatement. For young workers, unemployment means not only the loss of output and earnings but, more important, the missed opportunity for on-the-job training and experience. The very high unemployment rate among low-skilled youth is symptomatic of the more serious problem that the jobs available to them generally offer little opportunity for training or advancement. The social cost of youth unemployment thus depends very much on the contemplated alternative. If we judge the social cost of youth unemployment by the type of no-training jobs that are currently available, the cost is relatively low. But if employment with useful on-the-job training is a feasible alternative, the social cost of youth unemployment is substantially greater than the immediate loss of output.

III. Conclusion

In this short paper, I have emphasized two basic points. First, the private cost of unemployment varies substantially and is often extremely low. This low private cost is an important cause of the permanently high unemployment rate in the United States. Second, the social costs of unemployment must be judged by considering the specific policy by which a worker would be reemployed.

In selecting these points for emphasis, I have ignored many of the issues generally associated with measuring the costs of unemployment. In conclusion, however, I want to call attention to two further costs of a chronically high unemployment rate that are likely to be of great long-run importance.

If we do not change the structural causes of our high unemployment rate, we will face growing pressure to adopt the strategy of some European countries that supress unemployment by denying firms the right to lay off workers without government approval and by denying those workers who lose their jobs the right to decide where and when they will return to work. In addition, a chronically high unemployment rate will create strong pressures for expansionary macroeconomic policies that will serve only to exaccerbate inflation. The loss of freedom in labor markets and the increase in inflation throughout the economy would be an extremely high cost to bear for our failure to reform the incentives and eliminate the barriers that create our unemployment problems.

REFERENCES

M. Feldstein, "The Welfare Cost of Permanent Inflation and Optimal Short-Run Economic Policy," Nat. Bur. Econ. Res. work. paper no. 201, New York 1977.

Stabilization Goals: Balancing Inflation and Unemployment

By HENRY C. WALLICH*

Unemployment and inflation are grave social ills; both capable, unless resolved, of changing our economic and perhaps political system. Between the two ills, moreover, there is only a very limited tradeoff. In the longer run, there is no tradeoff; indeed, they may tend to move in the same direction, if not at exactly the same time.

This last proposition is easier to defend today than it was twenty and even ten years ago when I first tried to argue it. That period spans the life and some might say the death of the Phillips curve, probably the most important innovation in macroeconomics since Keynes. Today, the defense of the proposition that there is very little tradeoff—and that only transitory—between unemployment and inflation can fall back upon the theoretical framework surrounding the "natural rate of unemployment," upon "rational expectations," and on a growing body of empirical research. It can fall back also upon the experience of the last dozen years. This experience has refuted the formerly widespread view that accelerating inflation is unlikely to occur without a continuously declining unemployment rate and would do little real damage if it did.

If ever there existed a meaningful tradeoff, it rested on workers' and employers' expectation that higher inflation would soon be reversed. Once experience ceased to validate that expectation, money illusion was bound to vanish quickly. With money illusion dissipating, any tradeoff will occur only along a Phillips curve shifting nearly concomitantly with changes in the rate of inflation.

*Board of Governors of the Federal Reserve System. I am indebted to David Lindsey for many helpful comments, to Daniel E. Laufenberg for statistical and other assistance, and to numerous associates for criticism. Errors are mine.

Even a Phillips curve that is vertical in the long run does not adequately explain present high and apparently stubborn levels of both unemployment and inflation. For most industrial countries, unemployment today seems to be above what one might suppose to be its natural rate. Yet inflation has moved to extraordinarily high levels and is declining very slowly, if at all. Many special reasons can be adduced—oil price increases, food shortages, raw material scrambles, errors of monetary and fiscal policy, uncompetitive wage and price behavior, exchange rate fluctuations. However, I believe that a more systematic pattern is discernible.

Inflation and unemployment have moved up together because a short-run Phillips curve that shifts over time in response to variations in inflation rates implies realized tradeoffs that change in accordance with the stage of the business cycle. When the economy expands, the curve traced out by unemployment and inflation becomes steep—much inflation must be accepted for a given reduction in unemployment. When the economy contracts, the curve traced out becomes flat—little reduction in inflation is accomplished for a given rise in unemployment. Where previously we recognized downward inflexibility of the level of wages and prices, today we are beginning to recognize diminishing downward flexibility of the rate of wage and price increases. Movements on the short-run Phillips curve, in other words, are not reversible.

The upward zigzagging of inflation and unemployment has been aggravated by the stop-go character of anticyclical policy. In the United States as in various other countries, policy has moved back and forth between fighting inflation while ignoring mounting unemployment and fighting unemployment while ignoring mounting

Table 1—Periods of Increasing and Decreasing Rates of Inflation and Unemployment (Four-Quarter Moving Averages)

Rate of Inflation[a]		Rate of Unemployment	
Decreasing	Increasing	Decreasing	Increasing
1960(I)–1961(I)			1960(I)–1961(III)
2.1 0.6			5.4 6.5
	1961(I)–1962(IV)	1961(III)–1962(IV)	
	0.6 2.1	6.5 5.3	
1962(IV)–1964(I)			1962(IV)–1963(IV)
2.1 1.3			5.3 5.4
	1964(I)–1966(IV)	1963(IV)–1967(II)	
	1.3 3.7	5.4 3.6	
1966(IV)–1967(II)			1967(II)–1967(IV)
3.7 2.5			3.6 3.7
	1967(II)–1970(I)	1967(IV)–1969(II)	
	2.5 5.6	3.7 3.3	
1970(I)–1972(II)			1969(II)–1971(IV)
5.6 3.8			3.3 5.8
	1972(II)–1975(I)	1971(IV)–1973(IV)	
	3.8 11.1	5.8 4.7	
1975(I)–1976(IV)			1973(IV)–1975(IV)
11.1 4.7			4.7 8.3
	1976(IV)–1977(III)	1975(IV)–1977(III)	
	4.7 5.6	8.3 7.2	

[a]The rate of inflation is a four-quarter moving average of the annualized percent change in the *GNP* deflator.

inflation. It is only recently that a more moderate approach has gained ground, seeking to reduce both evils simultaneously.

The net result has been the tracing out of a positively sloping relation between unemployment and inflation. The rough contours of this path for the United States are visible in Table 1 showing periods of increasing and decreasing rates of unemployment and inflation. The cyclical movements outlined in Table 1 are the result in considerable degree of policy measures, even though the precise consequences of those measures may not always have been adequately foreseen. This at least seems true of the United States, although not necessarily of other countries, where cyclical fluctuations often are imported. Could it then be argued that if no measures ever were taken to halt inflation, unemployment would never have to rise?

This would be tantamount to saying that continuously accelerating inflation might be indefinitely sustainable. Historical evidence indicates that it is not, and that hyperinflation in any event produces recession and unemployment. Even in the absence of acceleration, with inflation simply fluctuating around a high level, mounting unemployment ultimately seems unavoidable on present evidence. The reason is that inflation has shown itself to be adverse to investment and hence threatens a mounting imbalance between capital stock and labor force. In the United States, the growth of the capital stock clearly has not kept pace with that of the labor force. Full employment, by almost any definition, today would require operating the economy at rates of capacity utilization far in excess of historically noninflationary limits.

Thus, there seem to be three causal sequences through which inflation ultimately raises rather than reduces unemployment; 1) policy measures designed to curb inflations; 2) acceleration toward hyperinflation in the absence of such less than accommodative policy measures; and

3) disincentives to investment and reduction in the capital stock relative to the labor force.

In recent years, the economics profession seems to have modified its evaluation of the relative welfare loss from inflation and unemployment. In other words, in economists' perception, the indifference curve relating inflation and unemployment became flatter as the Phillips curve became steeper. Estimates of the loss from inflation have been raised while those of the loss from unemployment have been lowered. In the higher estimate of the loss from inflation, abandonment of the assumption of perfectly anticipated inflation has played a role. This useful analytical tool, like other forms of perfect foresight, has no reliable counterpart in the real world. The evidence so far seems to indicate that inflation will not be correctly anticipated.

Moreover, even in a world where inflation is correctly anticipated, making the correct adjustment to inflation could prove to be very difficult. Governments, indeed, will make every effort to prevent correct adjustment by insisting on original cost depreciation, on capital gains taxes based on nominal rather than inflation corrected values, on tardy adjustment of tax brackets, on interest rate ceilings, and on treating the inflation premium in interest rates as if it were income or an expense item, just to name a few of the roadblocks thrown up on the highway to adjustment to inflation. Official statements of the need and intention to bring down inflation have a similar effect.

Inflation therefore does affect real variables—the level and distributions of income and wealth, relative prices, investment, growth, and employment.

Furthermore, even in the unlikely event that rational expectations were to lead to unbiased anticipations of inflation, this does not guarantee systematic avoidance of real effects. Markets and institutions may not permit wealthholders to obtain the inflation premia they would like to have. Borrowers may not want to pay a nominal rate equal to the real rate plus the inflation premium. Alternative assets may not be available that would provide an adequate inflation-adjusted return. In that case it does not help the wealthholder who correctly perceives future inflation to "demand" such a return. The same problem could occur in the labor market.

Moreover, the contracting parties—employers and employees, lenders and borrowers—may not feel completely sure of their expectations. Each may therefore want to charge their counterpart a risk premium. This means that supply and demand functions, adjusted for the respective risk premia, will be shifted inward and will intersect only at a lower volume of transactions than they otherwise would.

Finally, any change in the rate of inflation, even if correctly anticipated for the future, after it has become effective will leave a residue of old contracts that cannot be adjusted and that give rise to redistributive gains and losses. One has to mount to a dizzying level of abstraction to lose sight of these individual consequences of inflation.

Moreover it is fanciful to discuss inflation in terms of perfect anticipations, however qualified. The fact that the *U.S.* government issues thirty-year bonds callable only after twenty-five years at about 8 percent does not imply that the government expects twenty-five years of inflation at about 5 percent—it is simply a sensible act of risk diversification on the part of a debtor. If inflation were firmly expected by government or the private sector to continue at some constant rate, forces would come into being causing it to accelerate. That, I fear, are our prospects today. The essence of inflation is uncertainty.

This means that under conditions of inflation, the ordinary uncertainties attached to individuals' income and wealth are greatly increased. The variance or risk term in utility functions rises. Since the variability of inflation has been shown to be positively related to its rate, risk rises with inflation and utility falls.

While the costs of inflation have been accorded increasing weight in professional opinion, the opposite has been the case with respect to unemployment. The costs of unemployment generally have been

evaluated at two levels: the macro-economic loss of total output, and the micro-economic financial and morale loss to those experiencing unemployment. The macro-economic loss presumably exceeds the sum of the micro losses thanks to the various compensation schemes that redistribute the impact.

The perception of loss of potential output attributable to unemployment is being reduced by the shift that has been taking place in the definition of the full employment level of unemployment. At one time, a plausible definition of full employment seemed to be the equality of unemployment and job vacancies. Today, the natural rate of unemployment seems to find increasing acceptance as the measure of full employment. The latter definition obviously leads to a lower level of potential output and hence a lower loss attributable to a given level of unemployment.

Additional doubts can be raised, moreover, about the concept of "potential output" as such. It rests heavily upon arbitrary institutional limitations, such as the forty-hour week and mandatory retirement at age 65. Today we seem to be in the process, with a minimum of fanfare, of raising the retirement age limit. Should we recompute past potential and compute the loss from not removing the limit earlier? Some dissatisfaction with the eight-hour day, too, seems to be indicated by the heavy movement of women into the labor force, some of which may reflect dissatisfaction with the earnings that husbands bring home from their eight-hour stint. What would potential be, and how much output would we be losing, if the workweek were forty-two or forty-five hours?

At the micro level, too, the cost of unemployment is coming to be seen in a more measured perspective. A considerable part of unemployment today is viewed as search activity, that is, as voluntary. Although far from painless, the benefits from search must be weighted against the micro costs of unemployment. Insofar as job search leads to better matching of skills and jobs, it produces gains also at the macro level.

In addition the economic cost of unemployment to the unemployed individual is perceived to be less disastrous than it has often been presented to be. Much unemployment is that of secondary earners in a household. Unemployment compensation is more adequate. Together with food stamps, tax deductibility of the benefits, savings on transportation, on meals away from home, and on clothing, unemployment "income" may come close, in many cases, to offsetting the wages lost to the individual. Any induced extension of unemployed status must then also be viewed as voluntary.

The transient character of much unemployment also is more clearly perceived. "The unemployed," for the most part, are not a fixed group like "the aged," but more nearly like "the sick." The composition of the unemployed part of the labor force is more clearly seen: unemployment is much lower among heads of households and particularly married males than among women and particularly teenagers. This fact, incidentally, also limits potential output from a given unemployed labor force—during an expansion, markets for skilled labor will tighten faster than labor markets in general.

All told, unemployment in liberal discourse is losing its role as a successor to sex among the Victorians—as an utterly obscene and unacceptable part of the human condition.

If inflation were thought to be costly mainly because it causes unemployment, and if unemployment itself were judged to be less costly than had earlier been thought, the issue of balancing the two would lose much of its portentousness. Such a misconception could arise from defining the respective "costs" in too narrowly economic terms. There is more to an economic system than the production of *GNP*. Indeed it can be argued that the most significant impacts of unemployment and inflation fall outside the area of determination of income and wealth.

In particular it is easy to overrate the importance of any loss of aggregate income and growth resulting from the joint and several impacts of unemployment and

inflation. Income per capita has tended almost to double in each generation. Does anyone argue seriously that earlier generations were substantially less happy than ours? That the 1950's or 1920's were periods of widespread distress? That 100 years ago, at an income per capita about one-tenth of today's in real terms, American lived in misery?

There have been enormous gains, of course. But they have principally consisted in the elimination, or at least reduction, of extreme conditions of poverty, and hardly from major gains in the sense of well-being of the average household. Growth has brought satisfaction probably because it has given income receivers a rising rather than simply a higher living standard. And growth, of course, has not resolved the dissatisfaction arising from the iron law of rank: for everyone who gains satisfaction by rising in the scale of income, wealth, or other forms of prominence, there must be another who has lost satisfaction by moving down.

If unemployment and inflation were to continue at high levels, the principal individual and social welfare losses would not come from income foregone. Nor need they come from diminishing satisfaction through a slower rate of growth, since it is at least conceivable that growth might continue at about the rate of the past, albeit on a lower path. The principal loss, it would seem to me, would take the form of a lowered quality of life, in the form of heightened uncertainty, sharper social conflicts, great injury to some individual life patterns, and mounting hostility to the economic and political institutions that would be held responsible. Persistent unemployment and inflation are forms of pollution of the social environment.

Unemployment directly affects a relatively small number, but with considerable intensity. High turnover increases this number and softens the impact, as does improved compensation. However, in certain groups—not so much regional or occupational as age and racial, such as black teenagers—the condition with all its consequences is becoming endemic. Affected individuals and groups are in danger of moving outside the mainstream of society and becoming altogether alienated.

Little seems to be known about the consequences of this condition for the attitudes of those affected. A good deal has been said about the views and feelings of the unemployed, much of it derived primarily, one must assume, from introspection by overemployed economists. Given the paucity of objective studies, one may guess that plain hostility to the system must be at least as frequent a reaction as loss of personal dignity, frustration, and functional disturbances.

Inflation hits directly a much larger number than does unemployment, but generally far less severely and in many cases indeed with positive effects on income and wealth. Uncertainty, however, is bound to be pervasive under inflation. Partial indexation merely raises the risk of the unindexed remainder. The ability to provide for the future, an essential attribute of a civilized society, evaporates. Inflation, which early on had been thought to discourage saving, does nothing of the sort—in all major countries savings rates rose as inflation accelerated. Full protection of these savings can be offered only by government, to the extent it chooses to do so through indexation of social security, civil service pensions, and at some future point perhaps indexed bonds.

Some concluding remarks on policy seem in order. The standard prescription against inflation derived from the natural rate of unemployment analysis is to allow the unemployment rate to remain above the natural rate for some time. To the extent that, by design or default, this prescription has been employed, it has so far given poor results. This experience reflects the view expressed earlier, that when inflation is on the way down, the short-run Phillips curve becomes quite flat.

A type of action that would simultaneously reduce unemployment and inflation is the family of tax-oriented incomes policies (*TIP*). Numerous versions of *TIP* have been proposed. Restraint can be exerted through tax penalties, or

through tax bonuses. It can be exerted against wages only, on the well-validated assumption that prices are closely tied to wages, or against both wages and prices. Applicability can be compulsory or voluntary. The taxes used can be the corporate income tax, or a payroll or sales tax. The principle is always the same. There are no mandatory controls. Market forces continue to govern. If a firm wants to concede a high wage increase, for whatever reason, it is free to do so, provided the tax is paid. Only the balance of bargaining power is shifted in favor of restraint.

The principle of *TIP* is to internalize to the wage and price setter the inflationary externalities he creates. The effect would be to break into the present spiral in which inflation moves forward mainly by its own momentum. The result should be not only a decline of inflation, but also an opportunity for lower unemployment. The Phillips curve or, if one prefers, the natural rate of unemployment, would have been moved toward lower levels of unemployment.

There are difficulties to be overcome, both technical and political. In the light of the high social costs of unemployment and inflation, I regard the effort as eminently worthwhile. Those who do not share the view expressed here that these are the principal costs of those twin evils but are primarily concerned about their economic cost, or who continue to believe in the existence of a meaningful tradeoff between them, should find the proposal no less convincing.

EARNINGS AND EMPLOYMENT OF WOMEN AND RACIAL MINORITIES

The Structure of Female Wages

By MARY CORCORAN*

While most people would agree that familial responsibilities affect women's labor market behavior and wages, surprisingly little is known about how this process operates. Past investigations of women's wages have generally relied on data sets designed for other purposes, and, as a result, theoretically important determinants of women's wages may be measured imprecisely or omitted entirely from analyses.

Familial responsibilities influence women's labor market behavior in at least two distinct ways. First, many women will withdraw entirely from labor market activities to bear and/or raise children. Not only does this reduce the total amounts of work experience and job tenure women acquire, but Jacob Mincer and Solomon Polachek further argue that women's human capital (work skills) will depreciate during such withdrawals and that such withdrawals will affect the timing of women's investments in on-the-job training. Second, women who choose to work may adjust their labor market activities to meet family responsibilities in ways which reduce productivity and hence wages. For instance, women with home responsibilities might restrict job locations or schedules, or might take off extra time from work to care for sick children.

The 1976 Panel Study of Income Dynamics (*PSID*) is well suited for exploring female wages. The *PSID* is a longitudinal study of 5,000 families which began in 1968. In 1976, male heads of household, female heads of household, and wives were asked to provide detailed information on earnings, education, work history, absenteeism, and self-imposed restrictions on job hours and job location. These data are used to describe women's patterns of work history and labor force attachment, to specify the determinants of women's wages, and to investigate the wage gaps between white men and each group of women.

*Assistant professor of political science, University of Michigan. This paper draws on data analyses conducted by Richard Coe, Greg Duncan, Martha Hill, and Saul Hoffman as well as myself. I have received financial support from the Department of Health, Education, and Welfare and the Labor Department. Karen Mason has commented on an earlier draft of this paper. None of these individuals or institutions is responsible for opinions or errors in this paper.

I. Patterns of Work History

The majority of women withdraw from the labor market at some point in their careers. In the *PSID*, only 36 percent of employed white women 18 to 64 had worked continuously since school completion. This drops to 25 percent if we look at employed white women 30 to 45. White women varied considerably on when they timed labor force withdrawals. About 29 percent of employed white women 18 to 64 delayed beginning work after school completion (for an average of 9.6 years) and then worked continuously; another 15 percent worked a while, dropped out of work (for an average of 8.2 years) and then returned to work and worked continuously; still another 20 percent experienced two or more prolonged periods of nonwork after school completion.[1] Black women's work history patterns involved fewer interruptions once a work career had begun and

[1]The percentage of women with two or more periods of nonwork is greater than the percentage of women with two or more interruptions; because periods of nonwork can occur after school and prior to starting a work career as well as when a work career is interrupted.

TABLE 1—MEANS AND REGRESSION COEFFICIENTS FOR WORK HISTORY AND LABOR FORCE ATTACHMENT MEASURES FOR EMPLOYED WHITE WOMEN AND BLACK WOMEN AND WHITE MEN 18–64
Dependent Variable = *ln* (1975 hourly wage)

Independent Variables	White[a] Women		Black[a] Women		White[a] Men	
	$\bar{X}$	b (s.e)	$\bar{X}$	b (s.e)	$\bar{X}$	b (s.e)
Work History Measures[b]						
Years between school and work[c]	3.15	−.0076[h]	3.07	.0032	1.66	−.0053
		(.0021)		(.0031)		(.0097)
Length of most recent interruption (years)[d]	2.52	−.0004	0.538	.0017	.97	−.0029
		(.0025)		(.0077)		(.0070)
Interrupted two or more times[d]	0.118	.0121	.037	−.1186	.028	−.0331
		(.0392)		(.0897)		(.0668)
Experience prior to present job (years)	8.05	.0083[h]	9.27	.0103[h]	11.27	.0141[h]
		(.0042)		(.0052)		(.0034)
Experience prior to present job (squared)	129.9	−.0003[g]	161.9	−.0003	225.0	−.0003[h]
		(.0001)		(.0002)		(.0001)
Tenure in present job (years)	5.74	.0245[h]	6.45	.0142[h]	8.72	.0242[h]
		(.0020)		(.0025)		(.0013)
Proportion of total working years that were full time[e]	0.790	.2883[h]	0.826	.1620[h]	.909	.3365[h]
		(.0460)		(.0597)		(.0009)
Labor Force Attachment Measures[f]						
Whether placed limitations on job hours or location	0.342	−.0416	0.216	−.0056	0.145	−.0603[h]
		(.0268)		(.0394)		(.0303)
Annual hours of absenteeism due to own illness	43.38	−.0002[g]	58.01	.0003	36.50	.0002[h]
		(.0001)		(.0002)		(.0001)
Annual hours of absenteeism due to illness of other family members	12.45	−.0001	25.68	.0001	4.01	.0005
		(.0002)		(.0001)		(.0005)
Whether working part time voluntarily	0.148	.0438	.090	−.0448	.010	.0673
		(.0369)		(.0610)		(.1077)
Other Variables						
Formal education (in years)	12.73	.0848[h]	11.75	.0909[h]	12.85	.0638[h]
		(.0051)		(.0074)		(.0038)
Size of Largest City in Area (in 100,000's)	4.08	.0191[h]	5.26	.0179[h]	3.84	.0201[h]
		(.0030)		(.0042)		(.0026)
Whether residing in the South	0.261	−.0200	0.558	−.1283[h]	0.266	−.0528[h]
		(.0294)		(.0365)		(.0248)
N (Number of observations)	1326		741		2250	
$\bar{R}^2$		.312		.301		.287
1975 Hourly Wage (Geometric mean)	3.61		3.17		5.60	
ln (1975 Hourly Wage)	1.284		1.154		1.722	

[a]In these analyses "white" refers to racial categories other than black. Observations include all employed household heads and wives 18–64. "Employed" refers to individuals who worked 500 hours or more in 1975.
[b]These work history variables are described at greater length in the author.
[c]Years between school and work measures the time spent not working after school completion and prior to taking one's present job.
[d]Interruptions refer to labor force withdrawals which occur after one's work career has begun.
[e]This is the ratio of years of full-time work to total years of work.
[f]These labor force attachment measures are described more fully in Hill and Coe.
[g]These two variables become significant when the insignificant work continuity measures are dropped.
[h]Significant at the .05 level.

these labor force withdrawals were shorter. Employed black women 18 to 64 were more likely than employed white women 18 to 64 to have worked continuously (42 percent [illegible] percent), or to have delayed labor [illegible] entry and then worked continuously (42 percent vs. 29 percent). Correspondingly, less than 20 percent of all employed black women had ever dropped out of work once their work careers had begun, and black women were only half as likely as white women to have experienced two or

more extended spells of nonwork. Finally, when employed women who have ever withdrawn from the labor market are compared, black women's spells of nonwork are on the average about 2 years shorter than those of white women.

Three measures were constructed to capture the timing, frequency, and duration of women's labor force withdrawals (see Table 1, col. 1). Years between school and work measures the length of labor force withdrawals which occur after school completion and precede one's work career. Two "interruptions" variables refer to labor force withdrawals which occur *after* a career has begun. Some women also cut back work experience by working part time rather than full time. On the average, about 20 percent of the work experience of employed women 18 to 64 involved part-time rather than full-time work.

II. Labor Force Attachment

Women are believed to adjust labor market behavior and job choice so as to accommodate child rearing and family responsibilities, with such adjustments lowering productivity and hence wages. The *PSID* provides four direct measures of ways in which women might adjust labor market behavior: absenteeism due to own illness; absenteeism due to illness of others; self-imposed restrictions on work hours and/or job location; voluntary part-time work. In addition, the wage changes experienced by employed wives who moved between 1970 and 1975 will be compared to wage changes experienced by employed wives who did not move during this period. If wives moved in order to accommodate a spouse's career, this should negatively affect wages.

A considerable number of women appear to adjust their labor market behavior. About one-third of all white women and one-fifth of all black women reported that limitations on job hours and/or job location were factors in taking their present job. Interestingly both groups of women are more likely to report that the *timing* rather than the number of hours or job location was a factor in taking the present job, suggesting that policies which increase the flexibility of work hours might help working women to accommodate family needs. About 15 percent of white women and 9 percent of black women worked part time voluntarily in 1973.

About half of all employed women were absent from work in 1975 because of their own illness, and 20 to 25 percent were absent because of another family member's illness. Workers who missed work to care for other family members missed a fair amount of work: an average of 57 hours/year for white women and 103 hours/year for black women. And about 65 percent of all women who missed work to care for family members did so to care for sick children. Black women were absent from work much more frequently than white women. These black/white differences in absenteeism are not surprising given that blacks may suffer from poorer health than whites because of their generally more deprived backgrounds.

The labor force attachment measures included here are a substantial improvement over those available in past analyses. Some past studies, for instance, have interpreted the empirical effects of marital status and number of children on the relative earnings of men and women as indicators of male-female differences in labor force attachment. The *PSID* variables directly measure many of those behavior patterns resulting from the sex division of labor in the home, which have been expected to reduce women's wages.

III. The Determinants of Female Wages: Work History

Table 1, columns 3 and 5, show the results when *ln* (hourly wages) is regressed on the measures of work history, labor force attachment and schooling listed in column 1.

Contrary to theoretical expectations, women's work skills apparently depreciate only slightly, if at all, during periods of nonwork. Black women's wages were unaffected by labor force withdrawals. Labor

force withdrawals significantly reduced white women's wages only when such withdrawals preceded one's first job and followed school completion. Even then reductions in expected wages were small—less than 1 percent for each year out. Interruptions after a work career had begun had no effect on expected wages. Of course, the impact of withdrawals may be underestimated if some women use withdrawals to acquire and/or brush up on job-related skills. The *PSID* asked workers about schooling and skills acquired during work interruptions. Even when analysis is restricted to interruptions which involved neither schooling nor the acquisition of job-related skills, these interruptions still had no significant effects on women's wages (see the author).

If human capital depreciates during periods of nonwork, it is unclear why such capital should depreciate during one period and not another. One would expect, in fact, that the net rate of depreciation would be greater for interruptions in a work career than for delays in beginning work. Yet just the reverse is true for white women. One possible interpretation is that women who delay labor market entry are motivated differently than women who begin working directly after school. For instance, women who view themselves primarily as wives and mothers may be those most likely to delay starting work, and perhaps these motivational differences persist over time. On the other hand, entering the labor market directly after school completion may itself alter women's perceptions and motivations, so that women who work for a while after school completion may see themselves as potentially attached to the labor force throughout their lives.

These results differ from Mincer's and Polacheck's finding that labor force withdrawals had moderately large negative effects on wages in a national sample of married women aged 30 to 44 with children. But a replication of the Mincer-Polachek analysis with *PSID* data suggests that these apparent inconsistencies are caused by restricting the sample to women aged 30 to 44 years (see the author). Wages of women in this restricted age range appear to be quite sensitive to labor force withdrawals. This is consistent with the argument that women's wages upon return to the labor market are depressed for a short while because of misinformation about labor market opportunities. Other evidence supports this explanation. Women in the *PSID* were asked to report their hourly wages before and after their most recent interruption. On the average, white women's real hourly wages were about 19 percent lower immediately after labor market reentry than immediately before leaving.

Human capital theorists argue that wages increase with work experience because of training acquired while at work. Presumably then, those kinds of experience which involve more training will be more valuable than other kinds. A year of experience in one's present job, for instance, seems to be more valuable than a year of experience acquired in previous jobs (Table 1). And full-time work experience is clearly more valuable than part-time experience. A white woman, half of whose work experience has been in part-time jobs, for instance, earns about 14 percent less than an otherwise similarly qualified white woman, all of whose experience has been in full-time jobs. Mincer and Polachek argue that workers will invest more in training during periods following interruptions than in periods preceding interruptions. When I tested this in the *PSID*, I found that experience acquired prior to one's latest interruption was not significantly less valuable than later experience (see the author).

IV. The Determinants of Female Wages: Labor Force Attachment

The behavioral indicators of labor force attachment included in these analyses had negligible effects on women's wages. Women who were frequently absent from work or who had imposed limitations on work hours or job locations earned no less than did similarly qualified women who had attended work regularly and who had im-

posed no limitations. Similarly, women who were working part time voluntarily earned no less than other women. Finally, wage rate changes between 1970 and 1975 were computed separately for wives who moved during that period and for wives who didn't move (see Martha Hill). Differences in wage changes between the two groups were trivial. If variations in labor force attachment influenced worker productivity, this was not reflected in wages.

One might argue that labor force attachment is imprecisely measured because different kinds of self-imposed limitations are lumped together; those on the timing of job hours, on the amount of job hours and on job location. But when limitations were broken down into separate categories, limitations still never significantly reduced women's wages—with the single exception that white women who limited both job hours and job location earned about 13 percent less than other white women. Moreover, if we compare women to white men (Table 1, col. 2, 4, and 6), women consistently score lower than white men on labor force attachment measures—suggesting that these measures do capture sex differences in labor force attachment. Finally, excluded measures are probably correlated with our included attachment measures. Given the exceedingly weak effects of included measures, it seems unlikely that excluded measures will be important determinants of women's wages.

A second possibility is that employers tend to hire women into jobs where labor force attachment doesn't matter on the assumption that women are likely to be less motivated and less reliable than men. In this case men would be hired into jobs where labor force attachment is more important. If this were the case, women's wages should be less affected by labor force attachment than men's. This is generally *not* the case. If we compare the wage equations for white men to those of women, we see that white men's wages are no more affected by variations in labor force attachment than were either group of women's wages.

A third possibility is that this model misspecifies the relationship between labor force attachment and wages. For example, workers reported those job limitations which they imposed when they first took their present job. To the extent that wage effects of such limitations diminish over time, the equations reported in Table 1 will underestimate the effects of job limitations. But even when an interactive term between tenure and limitations is added to the women's wage equations, both the limitations measure and the interactive term are insignificant—suggesting that effects of limitations do *not* vary with job tenure (see Hill). A similar problem arises with the absenteeism measures. Absenteeism and wages are each reported for 1975. Employers may not yet have adjusted wages to reflect absenteeism for newly hired employees. But when interactive terms between tenure and the two absenteeism measures are added to the basic wage equation, results indicate that employers did not penalize tenured employees who missed considerable work time because of their own illness or because of the illness of other family members (see Richard Coe).

V. Earnings Differentials Between White Men and Women

In the *PSID* white men's hourly earnings for 1975 averaged \$6.67; this compares to averages of \$4.16 for white women and \$3.75 for black women. Some have attributed male-female wage differentials to the sex division of labor within the home: because women assume family and child-rearing duties, they restrict labor force activity in ways which reduce their wages relative to those of men.

We can investigate this explanation by assessing the extent to which male/female average differences on the labor force restrictions measured in Table 1 account for male/female differences in wages.[2]

[2]These figures are estimated by subtracting the female mean from the white male mean, multiplying the difference by the white male coefficient, and then expressing this product as a fraction of the differences in *ln* (hourly wages) between white men and women.

Average differences in work history, labor force attachment, and schooling account for 36 percent of the wage gap between white men and white women, and 27 percent of the wage gap between white men and black women. Much of this explanatory power is due to differences between white men and women in the amounts of experience, tenure, and proportion of full-time work experience. Differences in schooling also account for a considerable proportion of the wage gap between white men and black women. The substantial differences between the work continuity of white men and women (as measured by the first three variables in Table 1) accounted for less than 5 percent of the wage gap between white men and either group of women; and differences in labor force attachment accounted for *none* of the wage gaps between white men and either groups of women.

The overall explanatory power is hardly overwhelming given the extensive number of worker qualifications included in these analyses. While some relevant worker qualifications were likely omitted from this analysis, I suspect that at most half the wage gap between white men and women can be accounted for by average differences in worker qualifications. This strongly suggests that black and white women will not be paid the same as are white men with the similar qualifications.

VI. Conclusions

These analyses appear to support the following conclusions:

Women's labor market behavior resembles that predicted by cultural stereotypes. A majority of women have not worked continuously after school completion; much of women's work experience is part time; and even when women do work, a substantial minority place limits on when, how much, and/or where they will work and a substantial minority are absent from work to care for sick children.

Intermittency of work experience exerts almost no effect on women's wages; interruptions in work careers *never* significantly lowered wages for either black or white women; delays in starting work after school completion did lower white women's wages slightly.

Work experience, job tenure, and the extent to which past work has been full time all exert considerable impact on women's wages. This suggests that policy programs designed to increase women's economic opportunities should not be directed to programs at any single point in the career cycle but should instead be aimed at all phases of the career process.

Self-imposed restrictions on job hours or location, voluntary part-time work, absenteeism either for one's own illness or to care for others, and geographic mobility exerted no significant impact on women's wages.

Differences in work history, work continuity, labor force attachment, and formal schooling accounted for less than half the wage gap between white men and white women, and less then one-third the wage gap between white men and black women.

REFERENCES

R. Coe, "Absenteeism From Work," in Greg Duncan and James Morgan, eds., *Five Thousand American Families*, Ann Arbor 1978.

M. Corcoran, "Work Experience, Work Interruptions and Wages," in Greg Duncan and James Morgan, eds., *Five Thousand American Families*, Ann Arbor 1978.

M. Hill, "Self-Imposed Limitations on Work Schedule and Job Location," in Greg Duncan and James Morgan, eds., *Five Thousand American Families*, Ann Arbor 1978.

J. Mincer and S. Polachek, "Family Investments in Human Capital: Earnings of Women," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, Part II, S76-S108.

The Improving Economic Status of Black Americans

By JAMES P. SMITH*

While contemporary rhetoric often highlights differences between races, the data show that blacks are becoming less distinguishable from whites in at least one relevant index of performance—market earnings. Relative to white males, black male earnings have gradually increased, and the rise during the 1960's and the early 1970's is larger than that observed earlier. (See Table 1.) Yet, it is clearly the contrast between white and black females that is extraordinary. Twenty years ago the average black woman employed full time was earning approximately half the wage of a similarly employed white woman. By 1975, almost complete racial parity among women had been achieved.

In a recent article (1977), Finis Welch and I argued that the advance in the relative income of black males between 1960 and 1970 was due mainly to converging educational distributions by race and a narrowing in wage differentials between regions. Skill levels were relatively constant within cohorts and convergence was accomplished as increasingly similar racial cohorts entered labor markets while other less similar cohorts retired. Finally our test of affirmative action pressures indicated that before 1970 they had little impact. My first objective is to update our previous research to determine if the events of the last decade for males have continued unabated into the mid-1970's. Since a complete understanding of the dynamics of black-white changes necessitates explaining the patterns for females, my second goal is to expand the wage comparisons to include women.

The major explanations for narrowing in racial wage differences can be placed under four general categories. The central idea of the vintage hypothesis is that more recent black cohorts begin their job experiences with larger initial stocks of human capital, relative to whites, than previous cohorts. The second explanation involves migration. The rural-South to urban-North migration has partly been superceded by southern blacks moving to what are by now economically vibrant southern cities. The third category involves the effects of government affirmative action. Since 1970, it is alleged that a series of court cases imposing severe financial penalties on firms for noncompliance with affirmative action goals have added sharp teeth to government jaw boning. Finally, changes in other aspects of market work may be important in narrowing relative wages. This factor is more relevant for women than men and includes the choice of part- or full-time work, unique characteristics of certain occupations, and biases due to limiting comparisons solely to working women.

I. Evidence of Cohort Convergence

My principal explanation for the improved relative economic status of blacks is that successive cohorts of blacks and whites are simply becoming more alike in those attributes producing higher wages. In 1930, the average black new labor force entrant had credentials quite different from his white competitor. (See Table 2.) The typical black male (female) began a work career with 3.7 (2.6) fewer years of formal schooling than their white counterparts. Almost 70 percent of these blacks had a grade

*The Rand Corporation. This paper summarizes work from a large project on racial differences in income funded by grants from National Science Foundation and the Department of Health, Education, and Welfare. Part of this project is conducted with my colleague Finis Welch whose assistance on this paper is gratefully acknowledged.

TABLE 1—RATIOS OF MEDIAN WAGE INCOMES BY RACE AND SEX FOR SELECTED YEARS, 1947–75

	All Workers		Full-Time Workers	
Year	Black Females/ White Females	Black Males/ White Males	Black Females/ White Females	Black Males/ White Males
1947	.340	.543	.543	.640
1951	.421	.616	–	–
1955	.433	.588	.570	.635
1959	.532	.580	.664	.612
1963	.531	.568	.636	.654
1967	.643	.639	.765	.675
1969	.721	.666	.816	.694
1973	.896	.695	.882	.719
1974	.977	.709	.941	.736
1975	.973	.734	.986	.769

Source: Various issues of *Current Population Surveys*.

school diploma or less, and only 3 percent had any postsecondary schooling. As successive cohorts entered the labor force, the competitive disadvantage of blacks continuously dissipated. By 1970, 1.2 (.3) years of schooling separated black and white males (females) at the time of their initial labor force experience. Further, fewer than 10 percent of these new black workers had less than 9 years of schooling and almost a fifth had some postsecondary education. Even as current education levels by race converge, the weight of the past will continue to depress relative wages of blacks. However, in tracking changes over time, this generational improvement will contribute to convergence between races. Blacks not only have higher education levels relative to whites than their fathers did, but they also have parents with more education relative to whites than their parents had.

The story conveyed by nominal years of schooling is reinforced by data on schooling quality. The current, and often valid, criticism of the quality of contemporary black education makes us forget that the historical situation was much worse. The data (see Table 3) on nominal characteristics of schools tell a clear story of improving relative quality of black schools.[1] In 1920, black youths attended school only two-thirds as many days as white students, but there were no real black-white differences in days attended by 1954. Similarly, in 1920 teachers of black students had 1.75 as many pupils as the average teacher in the country. By 1954, this difference had

[1]For a detailed examination of this data, see Welch.

TABLE 2—YEARS OF SCHOOL COMPLETED AT ESTIMATED TIME OF LABOR MARKET ENTRY

	Year of Labor Market Entry									
	Males					Females				
	1930	1940	1950	1960	1970	1930	1940	1950	1960	1970
Mean Schooling of Blacks	5.9	8.0	9.9	11.1	11.4	8.0	10.1	11.2	11.6	12.1
Mean Schooling of Whites	9.6	11.1	12.0	12.6	12.6	10.6	11.4	11.9	12.0	12.5
Proportion of Blacks with less than 9 years of school	0.78	0.58	0.31	0.15	0.11	0.58	0.27	0.15	0.09	0.04
Proportion of Whites with less than 9 years of school	0.42	0.22	0.15	0.10	0.07	0.29	0.15	0.09	0.08	0.03

TABLE 3—COMPARISONS OF TWENTIETH CENTURY TRENDS IN CHARACTERISTICS BETWEEN THE SEGREGATED NEGRO SCHOOLS, SOUTHERN WHITE SCHOOLS, AND ALL U.S. SCHOOLS

	Average Days Attended Per Pupil Enrolled		Pupils Enrolled per Classroom Teacher		Enrollment in First Relative to Second Grade	
Year	Negro Schools	All Schools	Negro Schools	All Schools	Negro Schools	All Schools
1899–1900	57	69[a]	56.7	42.5[a]	1.37	1.14[a]
1908–09	71	88	56.4	39.9[a]	1.45	1.49[a]
1919–20	80	121	56.0	31.8	1.96	1.64
1929–30	97	143	43.7	30.0	2.35	1.48
1939–40	126	152	45.3	29.0	2.03	1.29
1949–50	148	158	33.6	27.5	1.62	1.20
1953–54	151	159	32.9	27.9	1.45	1.25

[a]Southern white schools only.

been substantially reduced. The extraordinarily high ratio of first to second graders suggests that on average a black student took about two years to complete the first grade in the 1930's. Retention rates that average 100 percent suggest low quality education coupled with inflexible standards. Between 1940 and 1954 implicit retention rates in southern Negro schools moved toward the national norm. It is difficult to link attributes of schools to measures of school achievement. Indeed, some of the measures offered may have had little impact on achievement, but the consistent picture of simultaneous convergence in all these dimensions makes the case for improving quality of black schools plausible.

II. Summary of Regression Results

Regressions for male wages were based on eight Current Population Surveys (*CPS*) for 1968–75 inclusive. In each year, separate samples were extracted for black and white males between the ages of 14 and 65. The dependent variable is the (*ln*) weekly wage and the explanatory variables fall into five classes: schooling, regional residence, market experience, direct and indirect government employment, and a set of estimated probabilities controlling for less than full-time work—the probability of zero earnings, part-time work, and full-time part-year work. A similar wage equation was estimated for black and white married women between the ages of 21 and 60 using the 1960 and 1970 U.S. Census. The dependent variable was again the (*ln*) hourly wage. The explanatory variables were divided into five classes: schooling, region, direct and indirect government employment, full- or part-time work, and age.

Using the ordinary least squares regression estimates, the change in the black-white wage ratio between 1960 and 1970 is

$$\Delta\, ln\, \bar{R} = [(\bar{X}_{B70} - \bar{X}_{B60}) - (\bar{X}_{W70} - \bar{X}_{W60})]b_0 + (\bar{X}_{B70} - \bar{X}_{B60})\delta_1 + (\bar{X}_{W60} - \bar{X}_{B60})\delta_2 - (\bar{X}_{B60} - \bar{X}_{W70})\delta_{12} - (\delta_{12}\bar{X}_{W70})$$

where b_0 is the parameter vector for white women in 1970, δ_1 is the black-white difference in parameters in 1970, δ_2 is the difference in white parameters between 1960 and 1970, and δ_{12} is the 1960–70 change in black-white difference in parameters. A similar procedure is used for changes in the male wage ratio between 1968 and 1975. The first term measures that part of the growth in black-white wage ratios due to contraction or expansion in black-white differences in characteristics. The other terms capture the impact of differential payments between races and across years. A detailed

description of the results is contained in the author, and the author and Welch (1977), and is only summarized here.

Almost one-half (one-third) of the rise in the ratio of black-white male (female) wages is explained by converging nominal characteristics. As new cohorts of black workers enter the market, the distinction between races in income-producing attributes is eroding and leading to a narrowing of the wage gap by race. Education plays the dominant role, explaining 47 (59) percent of the relative growth in black male (female) wages due to greater characteristic congruency. The large influx of black females into direct and indirect government employment explains an additional 20 percent of the female rise attributable to more similar attributes, while movement of black women from part-time to full-time jobs explains 12 percent. Migration accounts for only a small positive amount (9 percent) of the relative wage growth of black men and women.

The other terms in the equation provide additional insight into the differential structure of wage equations by race and the role of these differences in explaining improvement among blacks. I have estimated lower coefficients for elementary and secondary schooling for black males. If school systems are not an effective mechanism for increasing black male income, the problem is concentrated at the elementary and secondary levels. The marginal returns to college are actually higher for black males particularly in the early years of market experience, perhaps due to increasing black attendance at racially mixed northern colleges. In contrast to a declining white college coefficient, the returns to college for black males have remained stable over this period.

For all women, education coefficients rose dramatically with schooling level and increased over the decade. Female wages typically increased by less than 1 percent per year of elementary school education and less than 5 percent for the high school segment. No significant racial difference in education coefficients exists for women below the college level, but the premia to college are approximately 15 percent for black women and 10 percent for white.

The direct effect of converging schooling levels was partially offset by differential rates of return between races and across years. For example, in male comparisons the larger white schooling coefficients imply that whites gain more as schooling levels rise even if differences in schooling by race remain constant. Similarly, the increased return to female schooling over the decade benefited white women more than black women because they have more schooling.

In regional explanations of the changes in wage ratios the southern variables account for a significant part of the rise in both black male and female wages. Although migration flows had a small favorable impact on blacks, convergence towards the national norm in black-white southern wage ratios is far more important. Relative black-white wage ratios for both sexes have risen more rapidly in the South, especially among the young. For males with less than 10 years of experience and for women under 30 year old, relative black wages have increased by 10 percent more than in the rest of the country. For older workers, the male (female) ratio rose by 5 (3) percent more in the South. In addition to the general improvement for southern blacks, there was a reduction in black-white wage inequality among southern states and between urban and nonurban areas particularly among more recent cohorts. Vintage effects are greater in the South and affect all southerners, but they are larger among blacks than whites. New black southern workers apparently will enjoy career prospects that differ significantly from those which confronted their predecessors.

Direct and indirect government employment was used to test for the effect of affirmative action programs. Indirect government employment measures employment in industries regulated by either federal or state and local governments and the fraction of an industry's sales that go to either the federal or state and local governments. The argument is that if affirmative

action has an impact, it should be strongest on employment and wage trends in these industries. The public sector is becoming a more important employer of all blacks (relative to whites), but the changes observed for black males are small compared to those of black women. The proportion of black women employed in government rose from .18 to .28. For indirect government employment, the proportion of black women has also risen sharply both absolutely and relative to white females.

While the direct effect of increased government employment raised relative wages of black men and especially black women, black-white wage ratios have declined in these sectors. The large black wage gains were achieved in the private sector and not in those industries most susceptible to affirmative action pressures. In fact, this wage decline was so large that on balance the government variables actually predict a decline in the wages of black men relative to white men. The situation for black females is more ambiguous. The relative deterioration in black-white female wages in these government sectors also lowered the gains attributed to government, to a small positive amount. However, the magnitude of the employment inroads made by black women suggests that they are the most likely recipients of any beneficial effects of affirmative action. If quotas are imposed that include both race and sex as criteria, black women have a clear advantage. By filling two quotas for the price of one, they are the cheapest avenue open to employers to adhere to employment quotas. Any beneficial impact of quotas on groups that possess one targeted characteristic should be attenuated by the existence of a dual-attributed group.

The race-year intercept indicates that relative to whites the black regression line has shifted upwards over time. It is this shift combined with converging characteristics that explains the bulk of the rise in relative black income. This shift is the consequence of any secular improvement in the relative labor market value of black men and women not captured by nominal measures of characteristics included in earnings functions. While this improvement at the front end of the labor market is open to other interpretations, the evidence advanced earlier suggests that relative vintage effects for blacks deserve high priority. The age and experience variables indicate that cross-sectional black wage profiles have become flatter (relative to whites), so that vintage effects may in fact be accelerating over time.

III. Cohort and Life Cycle Comparisons: The Evidence for the Vintage Hypothesis

One feature common to all cross-sectional studies of black-white earnings differences is that younger blacks fare better in comparison to whites than their older counterparts. Secondary labor market theories of discrimination tended towards a life cycle explanation holding that over-the-career black earnings increase less rapidly than for whites. In the cohort view, however, the observed cross-sectional decline in relative black-white wages with experience simply reflects the fact that less experienced workers are simultaneously members of new cohorts. By comparing cross sections at different points in time, the potential of distinguishing life cycle and cohort effects is established.

For males, individual year *CPS* regressions are used to predict relative black-white wage paths with experience under two assumptions. The first is based on the earliest (1967) *CPS* regression and measures the predicted life cycle path from the cross section. The second uses the complete series of *CPS* cross sections and traces the predicted wage path for a given cohort in 1967 as it gains market experience. The 1967 cross-sectional patterns (see Table 4) uniformly predict declining black-white male wage ratios with years of market experience. It is this cross-sectional decline that gave credence to the secondary labor market view. However the within-cohort trends indicate that, if anything, black-white male wage ratios have increased over the career, especially for more schooled workers. In 1967, the cross

TABLE 4—COMPARISONS OF CROSS-SECTIONAL AND LIFE CYCLE BLACK-WHITE WAGE RATIOS

	A. Males Years of Market Experience				
	1	5	10	15	20
Schooling = 16					
1967 cross-sectional observation/prediction	.834	.806	.779	.760	
life cycle prediction		.892	.912	.842	.792
Schooling = 12					
1967 cross-sectional observation/prediction	.853	.817	.781	.755	
life cycle prediction		.852	.829	.799	.783
Schooling = 8					
1967 cross-sectional observation/prediction	.892	.817	.880	.808	
life cycle prediction		.913	.844	.819	.808

	B. Females Age			
Year	25–34	35–44	45–54	55–64
1967	.731	.753	.716	.584
1971	.924	.879	.837	.696
1975	.971	.926	.871	.734

section predicted that the wage ratio of those with 16 years schooling and 5 years experience would decline from .806 to .779 with five or more years in the market. However, by 1972 the predicted wage ratio for this cohort had increased to .912. Therefore, the weight of the evidence supports cohort improvement and rejects the secondary labor market hypothesis. There is also clear support for strong vintage effects for black females with the most rapid relative wage improvement accruing among the youngest females. There also exist substantial wage gains within cohorts. Factors that operate mainly at the front end of the labor market apparently are not the sole cause of these recent trends for females.

IV. Additional Reasons for the Rise in Black Female Wages

In 1960, over one-third of all married black working women were domestic servants in contrast to only 2 percent of white women. More than 25 percent of 21–25 year old black women were domestics, so that it was an important source of employment for new entrants. But over the decade, the fraction of black women in domestic services declined to 14 percent with less than 3 percent of the youngest black women employed there. The changes within the South were even more dramatic. Half of all employed southern black women were domestics in 1960, but this proportion fell to less than a quarter by 1970. Particularly noteworthy are the trends observed among younger workers where the proportion declined from 50 to 5 percent. Since nonpecuniary and nonreported wages are purported to be a large part of total compensation for domestics, the real extent of the relative wage improvement of black women may be overstated. Given the historical importance of domestic service for black women, an in-depth study of this market is obviously required before any complete understanding of the recent relative wage improvement of black women can be achieved.

Between 1960 and 1970, the percentage

of black women working less than thirty hours declined from .27 to .17. Blacks in 1960 were more likely to have part-time jobs than white women, but this reversed over the decade, suggesting that new white market participants have selected part-time work. Blacks gained in three ways from this shift into full-time jobs. Most directly, their observed weekly and annual earnings are higher as a result of their increased work effort. Part-time jobs also tend to be transitory over time, so that full-time work may signal a more permanent commitment to the labor force. Finally, if full-time employees receive higher wage rates, observed mean black wages will rise as they shift into full-time work. The problem is that my Census estimates indicate that in 1970 black women earned 8 percent less per hour when engaged in full-time work, so that the shift into full-time jobs predicts a fall in black-white female hourly wages. Because of the absence of direct information on hourly wages in the Census, it is clearly not the ideal data for estimating breakpoints in the wage-hours locus. More appropriate data will ultimately determine the importance of this adoption of full-time employment in raising black female wages.

One difficulty with wage comparisons among women is that wage rates are directly observable only for working women. This "selectivity" bias could distort measured average wage differences among groups of women differing in their labor force participation rates. When female wage equations were reestimated correcting for the selectivity bias, the preliminary results indicated that this bias may have contributed to the recorded rise in the relative wage of black women. Among white women, the average wage of working women exceeds that predicted for all white women by the regression, but the opposite appears true for black women. Thus nonworking black women would tend to receive higher wages than the currently employed black women, but they choose not to participate because their nonmarket opportunities are even greater. Thus one cause of the more rapid rise for black women is that as their participation rates rose over the decade, wages of the average working black female increased because the additional workers received wage offers that exceeded those available to those already working.

V. Conclusion

In general, the variables that explained the relative improvement of wages of black workers observed between 1960 and 1970 were also the principal reasons for the more recent improvement among black males: increased congruency in education and the narrowing of between-region racial wage differentials. Confirming a conclusion reached for the 1960's, affirmative action programs were not a major contributor to rising relative wages of black males. The increased similarity in education distributions and the rapid rise in black wages in the South were also important reasons for the remarkable wage advances of black women. While the evidence on affirmative action is mixed, the primary beneficiary may well have been black women. Adoption of full-time jobs, elimination of domestic services as their primary occupation, and the increase in participation rates in light of preliminary results on the relevant wage of new entrants have all contributed to the relative rise in black female wages. Finally, my evidence rejects the secondary labor market view that black males and females are relegated to dead-end jobs with little career growth potential. Rather, it favors the alternative vintage hypothesis that more recent cohorts of blacks are more similar to whites in marketable skills than were their black predecessors.

REFERENCES

J. P. Smith, "The Convergence to Racial Equality in Women's Wages," in Cynthia Lloyd, ed., *Women in the Labor Market*, forthcoming.

——— **and F. Welch**, "Black-White Male Ratios: 1960–70," *Amer. Econ. Rev.*,

June 1977, *67*, 323-38.

——— and ———, "Race Differences in Earnings: A Survey and New Evidence," in Peter Mieszkowski and Mahlon Straszheim, eds., *Issues in Urban Economics*, Vol. 2, forthcoming.

F. **Welch**, "Education and Racial Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton 1973.

U.S. Bureau of the Census, *Current Population Surveys*, various issues.

The Economic Status of Blacks and Whites

By MARCUS ALEXIS*

The civil rights activities of the 1950's and 1960's awakened interest in the social, political, and economic disadvantages under which large numbers of American blacks lived. There is no need to chronicle the segregation, discrimination, and denial of civil liberties and political freedoms. But a decade after the passage of the 1964 Civil Rights Act and the 1965 Voting Rights Act there is still a raging controversy over the "progress" that black Americans have made. Most blacks, while agreeing that opportunities were created by the activities of the 1950's and 1960's, are not as sanguine about the results and future prospects as are most whites. The major concern of this paper is to suggest an explanation for this apparent conflict in interpreting the statistics on black and white economic conditions and to suggest measures to resolve the controversy.

Different readings of the same economic data are not rare. The disagreement between trade unionists and employers regarding the relative importance of unemployment and inflation is a case in point. But, as will be suggested below, the differences in view which lead black leaders to emphasize the unfinished job before us and white leaders to emphasize the "improvements" which have taken place have as much to do with what is (or is not) measured as they do with the statistics used to measure the phenomenon.

The measures commonly employed in economics—income, wealth, wages, prices, unemployment rates, for example—have been compiled to assess the relative positions of blacks and whites. Ben Wattenberg and Richard Scammon, in an article that received considerable public attention, concluded that real progress had been made by blacks and that more progress was likely. This view was challenged by Herrington Bryce and Karl Gregory. In this case the debate was about the representativeness of the data and not (so much) about whether the right things are being measured.

Even when there is agreement about the accuracy and representativeness of the statistics used to measure the status of blacks and of whites, differences of opinion regarding interpretation might still arise. Consider for example the controversy over whether the relevant measure of black to white income is the ratio of the two incomes or the difference between them. During the 1960's the ratio of black to white increased, but so did the dollar gap between median black and median white income. Which is a better measure of relative position is a question which cannot be dismissed lightly. Simply using more complete income distributions, such as income at quintiles, as is done by Albert Wohlstetter and Sinclair Coleman does not resolve the problem. Thus another purpose of this paper is to suggest an alternative measure of relative positions, which measures both the difference (albeit an adjusted difference) and the ratio of the incomes.

I. Economic Status and Well-Being

The reason we are supposedly interested in data on economic conditions is because we believe there is a relation between economic conditions and "well-being." Economic resources place at our disposal command over a wide range of goods and services. These goods and services in turn serve the useful purpose of satisfying some

*Professor of economics and urban affairs, Northwestern University.

appetite (desire) we have. In the standard textbook treatment the goods and services that are consumed are market goods. Economic resources also can be used to purchase nonmarket goods (for example, peace of mind) or goods for which there are informal markets—influence, prestige, name recognition, and the like.

When blacks and whites are compared in terms of income, the implicit welfare measure is a bundle of goods and services purchasable with an income of a given size. If, for example, blacks are found to have an income that is three-fifths of white income, then blacks can consume only .6 times the goods and services available to whites. But at some point this ratio of available consumption approach looses its attractiveness. Suppose the income ratio of blacks (B) to whites (W), $B/W = 1/1{,}000$, with median black incomes equal to $1,000 and median white incomes equal to $1,000,000. Does anyone really believe that the millionaire consumes 1,000 times the *same* goods and services as the person with a $1,000 income?

As incomes rises, wealthy persons can consume more of the goods poorer persons consume but can also include in their consumption sets goods not available to poor consumers (due possibly to indivisibilities). Thus the wealthy can consume yachts, chateaux, original paintings of the masters, political influence, positions of honor (university trustees, titles, etc.), and similar goods unavailable to those with lower incomes. Goods and social and political opportunities are not continuous linear functions of income. Consumption sets available at some incomes include elements unavailable at lower ones.

II. Power and the Environment

In my 1973 article it was pointed out that power relations are at the heart of discrimination and that while economists focus on the marketplace, the importance of income and wealth is in the ways in which they give individuals and groups opportunities to determine the social, economic, and political environments.

In the traditional model of economic analysis of political behavior, four basic assumptions are made. First, we assume the presence of scarcity. Second, methodological individualism; the individual is the basic unit of analysis. Third, self-interest; individuals undertake actions because they stand to benefit from them. Fourth, individual rationality; in making choices an individual will choose the course that will give him (her) the greatest satisfaction—maximizing behavior.

In viewing how collections of individuals act together to achieve common objectives we eschew the individualistic assumption. This departure from the traditional mode of analysis has been applied in my (1973) article, James Buchanan, and Rubin Saposnik, among others. The author and Buchanan are concerned with externalities; in the former an interdependence of utility functions, and in the latter an externality in the consumption of a (quasi) public good.

Saposnik is concerned with the way in which particular economic environments are determined. The environment may be viewed as a vector of outcomes, each element relating to a particular variable subject to control—size of the public sector, unemployment rates, levels of inflation, black to white wage rates, and so forth. Power is defined as the ability of individuals or groups to influence the choice of environments by society. In each environment there is associated with an individual an index of power. The economic environment is not totally determined by individualistic market behavior; for example, the size of the public sector is one element of the environment.

In particular each individual is assumed to have available to him some power, measured by a power index $k_i(e)$. The preferences that society adopts correspond to the preferences of the collection of individuals in the (coalition) with the largest aggregate power index. In the case of two alternative environments $(e_1, e_2) \epsilon E$ there is defined a set valued function $G(e_1, e_2) =$

$\{i | e_1 R_i e_2\}$ and the social preference

$$K^e G(e_1, e_2) = \sum_{i \in G(e_1, e_2)} K_i(e)$$

where R_i is the preference ranking of the ith individual. Outcomes are clearly the composite of a set of individual preferences.

As usual, nothing is said about how these preferences are obtained. If, as in my 1973 article, it is assumed that members of some collection(s) perceive that their interests are related, then they will be more likely to coalesce. And if they have a larger aggregate power index, then they will obtain their preferences. Of course, one should not assume that the power index is determined by absolute numbers of individuals. A numerical minority can be a power majority in the sense that it has a larger power index, thus obtaining a favorable economic environment. South Africa presents such a case, where the white minority is a power majority.

There are incentives to have the winning coalition as small as possible (see William Riker). The reason is that the benefits are then divided over a smaller number of persons. This is of some importance in cases where some privilege is conferred by a particular environment. This preferred treatment may be in the form of taxation of income and/or wealth, wage rates (as in South Africa), occupational or residential choice (as in the United States), or any number of cases in which particular environments are associated with specific preferences.

There are sometimes costs involved in forming coalitions. These costs may be generally described as organization costs. The optimal size of the coalition will then depend on the marginal benefits of membership and the marginal costs of organization. Some desired outcomes require larger power coalitions than others. A simple majority will do to revise the Internal Revenue Code, but larger majorities in the Congress and the state legislatures are required for constitutional amendments abolishing the income tax or limiting the size of the public sector to some specified fraction of GNP.

III. Initial Conditions and Long-Run Equilibrium

Suppose one starts with a situation in which one group is at a disadvantage; they are never in the winning coalition and are therefore unable to choose environments favorable (or neutral) to them. The outcomes of the social process exercise leave them with lower wage rates, lower incomes, higher unemployment, unfavorable occupational distribution (more seasonal and cyclically sensitive), and smaller endowments of human capital. These are characteristics of the black population in the United States. Now suppose that the feasible environments include outcomes which do not discriminate against those in the losing coalition. Equal pay must be given for equal work, and workers with the same human capital endowments face the same employment opportunities. Will the elimination of the market discrimination lead to equality of outcomes in the long run? Specifically, will the incomes of blacks and white converge to a 1/1 ratio over time? This is the question addressed by Glenn Loury. He concludes that the answer is "not necessarily."

In Loury's analysis the assumptions are blacks and whites have 1) the same distribution of innate abilities, 2) individually identical utility functions for acquiring labor market characteristics (human capital), 3) different socioeconomic factors, and 4) different communities. Socioeconomic status is determined by race and parental income. Each person is completely characterized at birth by innate endowment, race, and parental income. The acquisition of productive characteristics is by a social process which is an interaction of home, community environment, and educational institution. Employment opportunities of adults are determined by the characteristics acquired through the social process in the preadult period.

Loury is able to show that under some

reasonable conditions black and white incomes will not converge. A sufficient condition for convergence is that the ratio of black to white income in period t + 1 is greater than it is in period t.

IV. Income and Power

The relative power index of an individual $k_i(e)/\Sigma K_i(e)$ is not merely his share of total income; it can be greater or less than that. Power flows from economic resources—income and wealth—and from ability to persuade others to behave in a manner so as to increase the chances that one's preferred outcome(s) are realized. The ability to persuade others to act in the desired manner might depend on other attributes of the individual, say his race, religion, kinship, fraternal ties (Masons, Elks, Knights of Columbus, etc.), or social contacts (alumni groups, Rotary, Lions, etc.).

Owing to the size of the public sector, attention must be paid to ways in which bureaucrats and elected officials act to increase the likelihood that the preferred outcomes of some individual(s) or group(s) are more likely to be realized. The recent successes of the "right-to-life" groups in withholding government funds from elective abortions can be seen as an example of one group producing a more preferred e_i through governmental intervention. In this case, a less powerful group (the poor) was forced to accept a change in the environment which is not binding on more affluent groups. This is an example of a group which would face a different set of market opportunities if their incomes were to increase by a few thousand dollars.

The affluent in our society have very different opportunity sets than the nonaffluent. They, almost exclusively as individuals, are the contributors to political campaigns, charitable and philanthropic causes, and to campaigns designed to influence the opinions of others. It would be surprising if the elected officials to whose campaigns they contribute, and the churches, colleges, community fund drives, hospital capital drives, and lobbying efforts to which they contribute are not responsive to their preferences. Indeed, some elected officials are said to be known as "friends" of special interest groups—organized labor or big business. Poverty and minority groups do not appear to have such steady supporters and are dependent on "goodwill."

It was suggested earlier that the relationship between income and power (as defined here) is not linear. How then does one relate income and power? And, equally important, what is the proper measure of income? Wohlstetter and Coleman argue for using the entire income distribution to evaluate changes in relative status. If comparisons are to be made at the medians, they argue that the ratio of the medians is to be preferred to the algebraic difference. It is reasonable to argue, however, that consumption opportunities are best measured by taking dollar incomes into account and the difference in the (median) dollar incomes is a better measure of inequality. The method used is not without some potential policy importance. While the ratio of black (actually blacks and others in Table 1) family income to white family income was increasing from 0.55 (1960) to 0.64 (1970), the dollar gap increased from $2,602 to $3,720. Thus while there was an 18 percent increase in the relative income of blacks there was also a 43 percent increase in the dollar gap. Wattenberg and Scammon focus on the former. Black leaders are more sensitive to the latter. The disagreement between the two on the extent of black economic progress is in part a reflection of this difference of emphasis.

V. An Appropriate Measure

A desirable measure of relative income 1) takes into account the income required for a minimum (subsistence) budget, 2) relates income differences to differences in consumption opportunities available, and 3) relates differences in consumption opportunities to individual (group) welfare. Income required for minimum consumption is nondiscretionary. Consumption opportu-

TABLE 1—MEDIAN INCOME OF BLACK AND WHITE FAMILIES: 1950–74 (IN CURRENT DOLLARS)

Year	Race of head			Ratio: Black and other races to white	Ratio: Black to white
	Black and other races	Black	White		
1950	$1,869	-	$3,445	0.54	-
1951	2,032	-	3,859	0.53	-
1952	2,338	-	4,114	0.57	-
1953	2,461	-	4,392	0.56	-
1954	2,410	-	4,339	0.56	-
1955	2,549	-	4,605	0.55	-
1956	2,628	-	4,993	0.53	-
1957	2,764	-	5,166	0.54	-
1958	2,711	-	5,300	0.51	-
1959	3,161	$3,047	5,893	0.54	0.52
1960	3,233	-	5,835	0.55	-
1961	3,191	-	5,981	0.53	-
1962	3,330	-	6,237	0.53	-
1963	3,465	-	6,548	0.53	-
1964	3,839	3,724	6,858	0.56	0.54
1965	3,994	3,886	7,251	0.55	0.54
1966	4,674	4,507	7,792	0.60	0.58
1967	5,094	4,875	8,234	0.62	0.59
1968	5,590	5,360	8,937	0.63	0.60
1969	6,191	5,999	9,794	0.63	0.61
1970	6,516	6,279	10,236	0.64	0.61
1971	6,714	6,440	10,672	0.63	0.60
1972	7,106	6,864	11,549	0.62	0.59
1973	7,596	7,269	12,595	0.60	0.58
1974					
United States	8,265	7,808	13,356	0.62	0.58
South	6,805	6,730	12,050	0.56	0.56
North and West	10,039	9,271	13,906	0.72	0.67
Northeast	9,399	8,788	14,164	0.66	0.62
North Central	9,901	9,846	14,017	0.71	0.70
West	11,107	8,585	13,339	0.83	0.64

Source: Current Population Reports, Table 9, p. 25. For definitions, see the source.

nities increase as the discretionary component of income increases. This suggests that income should be decomposed into discretionary and nondiscretionary components. A measure can then be constructed of discretionary income to total income, Y_d/Y.

For simplicity, let us take as the ratio of black to white incomes the percent that prevailed in 1973, 0.60. Also assume that the cost of a budget which satisfies consumption of material goods only, but no expenditures for power enhancing "goods," to be $6,000. It should be noted that this $6,000 "minimum budget" does not correspond to what is usually meant by a subsistence budget; in practice the goods only budget is much higher, perhaps $15,000 after taxes for a family of four at present day prices. If median black income is $6,000 and median white $10,000 (roughly the 1973 levels), then black families have a zero discretionary income/total income ratio and whites one of 0.40; the ratio of the measure for blacks and whites, $(Y_{db}/Y_B) \div (Y_{dw}/Y_W)$ is also zero. This means that blacks' incomes do not provide any resources to spend on power enhancement. Consequently all such private sector expenditures are made by whites. If the public sector does not contribute any outlays on behalf of either

blacks or whites, whites will have a positive power index and blacks one of zero.

As black incomes rise, opportunities for expenditures on power enhancing goods rise and there are opportunities to acquire power over the environment. When median black income is $9,000 and median white income is $15,000, the ratio of black to white income is maintained at 0.60 and with $6,000 the consumption of material goods only level, the ratio of the measure for blacks and whites is 0.56, only slightly lower than the ratio of black to white median income. And, when the income of blacks is $12,000 and whites is $20,000, the ratio increases to .71. Thus the opportunities for blacks to increase their share of power increases while the income ratio is fixed and the dollar gap increases. Both the relative income analysts and the dollar gap analysts would be understating changes in the relative power enhancing opportunities for blacks. Thus one interpretation of black critics of Wattenberg and Scammon is that at the low levels of present black incomes power enhancement is not relevant and that the deficiency in consumption of material goods is important. Furthermore, there is no income for creation of "social capital." These are valid points.

The approach suggested here considers issues which have not been discussed in other debates on the relative positions of blacks and whites. As black incomes rise, even if income ratios do not, blacks (at least more of them) will be in positions to apply their discretionary incomes in ways which will increase the relative influence of blacks on the environment. The abilities of blacks to support institutions of particular importance to them—churches, black colleges, foundations, etc.—rises with discretionary income. The possibility that black philanthropy (already underway in some cities with a Black United Fund) will support a major share of the financial requirements of such institutions suggests a dramatic change in the issues these institutions address. Dependence on government or white philanthropy has doubtlessly had an inhibiting influence. The same is true of black support of candidates for elective office. A large share of the campaign funds of black elected officials now comes from whites; the implications are obvious.

These possibilities are real. Even now, an increase in the income available for expenditure on recreation and leisure has led to a dramatic increase in the output of black oriented drama, dance, and literature. Black audiences are now large enough to support serious artistic ventures. With continued increases in income not required for material goods, blacks can have an important impact on the options available to them. In Loury it is argued that the communities in which blacks live have an important effect on the opportunities available to the offspring. What is argued here is that the opportunities to create social capital in these communities need not wait for equality of money incomes. Blacks may be closer to "self-determination" than has been previously believed.

VI. Relative Discretionary Income and Welfare

I made the point earlier that as income rises, other goods are now purchasable, and that some of these goods are not material market goods but are instead influence, status, pride, and the like. Additional discretionary income permits one to buy more of the material goods previously purchased and/or better qualities of the same goods (a larger, more aesthetic home), and makes available previously unattainable goods. Some of these latter goods give the purchaser greater voice in choice of environments and hence enhances the likelihood that the income and set of goods (and the prices at which they are available) correspond more to his (her) preferences. This necessarily increases welfare. Accordingly, increases in the ratio of black discretionary income/total income to that of whites increase black welfare. To know whether blacks have indeed improved their relative position we must know absolute incomes, and nondiscretionary consumption requirements. Neither the ratio of the two incomes nor the dollar gap alone is as sensitive an indicator of change in relative welfare.

VII. Conclusion

I have suggested a method for measuring the relative change in well-being of blacks and whites which is more sensitive than the ratios of (median) incomes or dollar difference in incomes and which requires one additional bit of information (albeit an important one), the level of material goods only consumption requirements. When this method is applied, we can arrive at conclusions similar to the conventional wisdom applied by many black leaders. Namely, an improvement in the relative income ratios is not sufficient evidence that the relative position of blacks has improved. We need to know what is happening to both total income and to discretionary income. The measure I suggest is the ratio of discretionary income/total income of blacks to whites, where discretionary is meant to denote that income at which power-enhancing consumption begins. An obvious refinement would entail calculating discretionary incomes for subclasses of blacks and whites, multiplying by the numbers of households and summing. The ratios of the resulting values would be superior to merely calculating discretionary incomes from the medians of the income distributions.

REFERENCES

M. Alexis, "A Theory of Labor Market Discrimination With Interdependent Utilities," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 296–302.

———, "Wealth Accumulation of Black and White Families: The Empirical Evidence: Comment," *J. Finan.*, May 1971, *26*, 458–65.

J. M. Buchanan, "An Economic Theory of Clubs," *Economica*, Feb. 1965, *32*, 1–14.

H. Bryce, "On the Progress of Blacks—A Comment on Wattenberg and Scammon," Joint Center for Political Studies, 1973.

K. Gregory, "Brief Report of The State of The Black Economy, 1973," *Rev. Black Polit. Econ.*, 1973, *4*, 3–15.

G. C. Loury, "A Dynamic Theory of Real Income Differences," in Phyllis A. Wallace and Annette M. LaMond, eds., *Women, Minorities and Employment Discrimination*, Lexington 1977, 153–86.

William Riker, *The Theory of Political Coalitions*, New Haven 1962.

R. Saposnik, "Power, the Economic Environment, and Social Choice," *Econometrica*, May 1974, *42*, 461–70.

B. J. Wattenberg and R. M. Scammon, "Black Progress and Liberal Rhetoric," *Commentary*, April 1973, 35–44.

A. Wohlstetter and S. Coleman, "Race Differences in Income," in Anthony Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972, 3–81.

U.S. Department of Commerce, *Current Population Reports*, Special Studies Series P-23, no. 54, *The Social and Economic Statics of the Black Population in the United States*, Washington 1974.

Discrimination in Mortgage Lending

By HAROLD BLACK, ROBERT L. SCHWEITZER, AND LEWIS MANDELL*

This paper is based on the Comptroller of the Currency-*FDIC* nationwide survey of banks that participated in a study of housing-related lending. It addresses two specific questions: what economic criteria are important in the banks' lending decisions; and do demographic variables such as race and sex appear important in the loan decision?

The sample of banks was stratified by the fourteen national bank regions, by whether the bank was inside or outside a Standard Metropolitan Statistical Area (*SMSA*), by size, and by the banks' mortgage loan activity within each stratum. Because the number of mortgages per bank was not known, consumer mortgage debt deflated by total loans was used as a proxy. That, of course, is an imperfect measure given that the size of mortgages vary. Nevertheless, it is felt that the mortgage/loan ratio gives some indication of a bank's mortgage activity.

The banks were mailed forms to be used in conjunction with every application for a home mortgage or home improvement loan of more than $4,000. The forms consisted of two parts. Part I, completed by the bank and mailed to the *FDIC* when a final loan decision was reached, requested information regarding the characteristics of the loan applied for as well as the applicant's financial position. Part II, completed and mailed to the *FDIC* by the applicant, contained information regarding personal characteristics of the applicant and coapplicant, if any.

The structure of the forms was intended to specify models that would explain the lending practices of the financial institutions. To this end, bankers were asked to comment on the relevancy of the items included on a draft of the bank form prior to the initiation of the survey.

The bank was requested to utilize the form for each in-person inquiry regarding a housing-related loan. The form allowed for the rejection of an application without collection of financial data if such loans were not being made at the time of the inquiry. However, the applicant form was still to be completed to insure that the bank's lending policies were not being inconsistently reported to the applicants.

Each form contained a unique identification number which allowed the two parts to be matched and the bank identified. The data were edited by the *FDIC* for completeness and consistency. Incomplete and/or inconsistent forms were repaired where possible.

I. Discrimination vs. Redlining

For the purposes of this paper, discrimination in the mortgage lending decision is defined as a differential action taken by one party which affects a second party based on the personal characteristics of the second party. This paper does not examine redlining which refers to differential action in real estate financing based on the geographic location of the property. Discrimination occurs if a loan application is rejected due to personal rather than economic characteristics of the borrower, while if redlining oc-

*Black and Schweitzer are deputy director and senior financial economist, respectively, Division of Economic Research and Analysis, Office of the Comptroller of the Currency. Mandell is professor and director, Division of Economics, Finance, and Business Law, University of Texas-San Antonio. David Dale, Norman Hannah, Daniel Williams, and Margaret Smith provided programming and statistical competence. William A. Longbrake, W. Roger Watson, and Paul Toxie of the *FDIC* facilitated editing the data set. David F. Rush, David A. Walker, and Robert R. Dince commented on an earlier version. The views herein are our own and do not necessarily represent those of the Office of the Comptroller of the Currency. Quotation of any of this material should include that disclaimer.

curred, it is conceivable that the applicant would be rejected solely because of the location of the property, *ceteris paribus*.

Discrimination may occur either by application denial or by approval of the loan with relatively unfavorable terms. Terms, however, are interdependent, for down payment percentage and interest rate are likely to be inversely related. Consequently, terms of loan should be estimated simultaneously to test for possible discrimination. A study of differences in loan terms by personal characteristics in Rochester, New York, by the New York State Banking Department found little evidence of discrimination but did not employ simultaneous estimation procedures (see Section V of that study).

This study investigates the issue of discrimination in the disposition of primary home mortgages. Two additional studies are scheduled for future completion, one on home improvement loans and the other on redlining. Both will utilize the data from the survey.

II. Survey Response

The survey instruments were mailed to the sample banks in September 1976. Three hundred institutions were selected: 82 national banks, 200 *FDIC* regulated banks, and 18 mutual savings banks. Of these, 176 chose to participate in the survey while 124 did not. However, 31 of the 124 banks that did not return the bank portion of the survey instrument had customers who returned the applicant portion to the *FDIC*.

Data collection ended at all institutions on February 15, 1977. In general, the form mailed by the applicant was received within a week of the application date. The bank's part with the final loan decision was received about six weeks later. In all, 13,613 bank forms were received, of which 12,079 survived the edit procedures. Also, 10,287 applicant forms were received of which 5,525 matched with the bank portion and 4,762 did not.

Analysis of the unmatched applicant forms shows that 273 were returned with no information regarding personal characteristics. The breakdown of the remaining forms is available upon request and shows no readily discernable patterns. There is no indication that banks systematically avoided returning forms that were matched with a group considered a priori as being most likely to suffer discrimination.

Table 1 shows the disposition of the matched forms by applicant characteristics. Of the 5,525 matching forms, 312 contained no demographic information and are excluded here. In 106 cases, disposition could not be ascertained. These are also excluded. The remaining 5,107 cases are described in the table. There are only 138 (2.7 percent) rejections.

Several reasons have been postulated for that rather low rejection rate. First, bankers may lend in the mortgage market only as a convenience for their better customers. That would imply that the rejection rate should be low at banks. Applicants who do not have the financial wherewithal would apply to savings and loan associations, whose role in the economy is to provide mortgage financing. If that were the case, one would anticipate that the rejection rates of banks should be significantly lower than those of savings and loans. Second, the banks could be prescreening applicants. Although the structure of the form was intended to consider prescreening, there is no assurance that this was successful. Third, real estate brokers or mortgage bankers could be prescreening applicants. They are not likely to seek financing for their customers at institutions which present a significant likelihood of rejection. Fourth, the survey itself could have been a factor. The possibility exists that the banks rejected fewer applications than otherwise for fear of possible actions by the regulators.

Table 1 shows acceptance/rejection rates by marital, racial, age, and sex categories. Bivariate analysis is utilized here for descriptive purposes only. Relevant economic variables that are crucial to the lending decision are not included in the table, therefore, no conclusions about possible discrimination can be drawn. The table shows little difference in the acceptance/re-

TABLE 1—NATIONAL DISPOSITION OF MATCHING APPLICATIONS BY APPLICANT CHARACTERISTICS

Characteristics	Approved and Accepted by Applicant	Approved and Withdrawn by Applicant	Rejected by Bank	Withdrawn by Applicant
All	4895 (95.9)	20 (0.4)	138 (2.7)	54 (1.1)
Sex				
Female	522 (95.1)	5 (0.9)	17 (3.1)	5 (0.9)
Male	4373 (95.9)	15 (0.3)	121 (2.7)	49 (1.1)
Marital Status				
Married	4011 (96.2)	11 (0.3)	109 (2.6)	37 (0.9)
Separated	59 (93.7)	0 (0.0)	1 (1.6)	3 (4.8)
Divorced	330 (96.2)	2 (0.6)	8 (2.3)	3 (0.9)
Widowed	108 (93.1)	2 (1.7)	3 (2.6)	8 (1.9)
Never Married	387 (92.8)	5 (1.2)	17 (4.1)	8 (1.9)
Race				
White	4436 (96.1)	18 (0.4)	115 (2.5)	46 (1.0)
Black	142 (87.1)	1 (0.6)	16 (9.8)	4 (2.4)
All Other	317 (96.4)	1 (0.3)	7 (2.1)	4 (1.2)
Age				
Under 20	10 (83.3)	0 (0.0)	2 (16.7)	0 (0.0)
20–29	1362 (95.2)	7 (0.5)	40 (2.8)	21 (1.5)
30–39	1635 (96.0)	6 (0.4)	45 (2.6)	17 (1.0)
40–49	998 (97.0)	4 (0.4)	20 (1.9)	7 (0.7)
50–59	672 (95.6)	1 (0.1)	23 (3.3)	7 (1.0)
60–69	178 (95.7)	0 (0.0)	7 (3.8)	1 (0.5)
70 and older	40 (90.9)	2 (4.5)	1 (2.3)	1 (2.3)

Note: The 106 cases for whom disposition was not ascertained are excluded. Percentages are in parentheses.

jection rate for male and female applicants. Marital status shows that those individuals who have never married have higher rejection rates than those who are married or have previously been married. The racial variable shows that blacks are more likely to be rejected than are other racial groups. Racial groups listed on the survey instrument were white, black, Puerto Rican, Mexican American, other Hispanic, Oriental, American Indian, and Other. The rejection rates for each were considerably lower than that for blacks. Last, the age variable appears to conform to the life cycle hypothesis with acceptances first rising as age increases, and then falling.

III. Discrimination Results

Linear models are employed to analyze the accept/reject decision of the participating banks. Because the dependent variable (accept/reject) is a binary variable coded one for reject and zero for accept, regression analysis is inappropriate. Rather, probit analysis provides a suitable statistical estimation procedure. Statistical significance levels of the coefficients are obtained and the significance of the model specification can be determined when the probability of limit and nonlimit responses is examined. A review of probit models is found in F. D. Nelson, p. 503.

The data compiled in the survey are partitioned into subsets. The first contains information regarding the characteristics of the loan. Included are the amount of loan, down payment, loan origination fees, years to maturity, simple annual interest rate, monthly loan payment, and loan insurance status. The second subset contains financial information on the applicant and property characteristics. The variables are employment, income, and net worth, as well as price, age, and appraised value of the property. Included within employment are years employed in current line of work and self-employed. Demographic information

from the applicant forms provides another subset including the applicant's age, sex, marital status, and race. The binary independent variables are coded as follows: insurance is one for insured loans and zero otherwise; self-employed is one if not self-employed and zero if self-employed; age is one if applicant is 55 years old or older and zero otherwise; sex is one for male applicants and zero for female applicants; marital status is one for not married and zero for married; and race is one for black applicants and zero otherwise.

Three models were estimated to analyze the probability of the accept/reject loan decision. In the first model, only the terms of the loan were included. The second model includes the economic variables. In the third model, personal characteristics are specified to test for discrimination.

Each variable is postulated as having a particular sign, a priori. Because the dependent variable is coded zero for accept and one for reject, those relationships are hypothesized as follows. A positive coefficient in the model implies that as the value of the independent variable increases, the value of the dependent variable approaches one (rejection of the loan).

As the amount of the loan requested increases, the probability of default rises, *ceteris paribus*, increasing the probability of rejection. This suggests that amount requested should have a positive coefficient. The same is asserted for years to maturity. Interest rate is assumed to have a positive sign associating lower interest rates with lower risk. Down payment, loan origination fee, and monthly loan payment are postulated as having negative signs. As the value of these variables decreases toward zero, the probability of acceptance decreases. Insurance is expected to be negative, for insured loans are postulated to be less risky than noninsured loans.

The results of the estimation of the three probit models are given in Table 2. All three of the models appear to be significant in explaining the bank's loan disposition decision. That is shown by the degree of significance of minus 2 times the *log* likelihood ratio. This summary statistic is the appropriate measure of goodness of fit for a probit model and has a *chi*-square distribution.

The results of the estimation of the first model are given in the first column of the table. Only the coefficients of down payment and interest rate are statistically significant. Both have the postulated sign and are significant at the 95 percent level. Because of possible geographic differences in mortgage markets, the model was estimated for two regions that had sufficient observations for analysis with no difference in results.

The economic variables, with the exception of monthly debt, are hypothesized to have negative signs while the age of house variable is postulated as having a positive sign. The second column in Table 2 shows that only self-employed is statistically significant among the economic variables. Age of house is also significant at the 95 percent level of confidence. Thus, as age of house increases, the probability increases that the loan application will be rejected by the bank. Note, however, that the economic life of the property and its current physical condition are not considered here.

The final model in Table 2 includes the personal characteristics of the applicant. Here the signs are hypothesized to be positive for marital status, race, and age. The hypothesis is that males, married people, nonblacks, and younger people are considered as being less likely to be discriminated against than are females, unmarried people, blacks, and older people. The nonblack category includes whites and all other applicant categories listed in Table 1. The estimated model shows that race is an important determinant in the loan decision at the 90 percent level of confidence. That is, blacks are less likely to be granted loans than are nonblacks, *ceteris paribus*.

Note, however, that the summary statistics show a large improvement between Models 1 and 2. The change between Models 2 and 3 is not large. Therefore, although the racial variable is statistically significant at the 90 percent level of confidence, one must interpret its overall

TABLE 2—DETERMINANTS OF ACCEPTANCE OF HOME MORTGAGE APPLICATIONS: ACCEPT = 0, REJECT = 1 (PROBIT ANALYSIS)

Independent Variables (Postulated Sign of Coefficient)	Model 1	Model 2	Model 3
Terms of Loan			
Amount Request	0.7448×10^{-6}	0.3135×10^{-5}	0.2695×10^{-5}
(+)	(0.16)	(0.65)	(0.55)
Downpayment	-0.2154×10^{-4}	-0.2205×10^{-4}	-0.2147×10^{-4}
(−)	(−3.18)[a]	(−3.12)[a]	(−3.03)[a]
Loan Origination Fee	-0.1592×10^{-3}	-0.1798×10^{-3}	-0.1786×10^{-3}
(−)	(−0.75)	(−0.84)	(−0.84)
Years to Maturity	0.3385×10^{-2}	0.5249×10^{-2}	0.5269×10^{-2}
(+)	(0.37)	(0.56)	(0.56)
Interest Rate	0.1951×10^{-2}	0.2050×10^{-2}	0.1935×10^{-2}
(+)	(2.68)[a]	(2.77)[a]	(2.56)[a]
Monthly Payment	0.2159×10^{-3}	0.1871×10^{-3}	0.1836×10^{-3}
(+)	(0.99)	(0.84)	(0.82)
Insurance Status	0.0367	0.0316	0.0317
(−)	(0.35)	(0.28)	(0.30)
Economic Variables			
Total Income	—	0.5315×10^{-5}	0.5200×10^{-5}
(−)		(0.49)	(0.48)
Net Worth	—	-0.7999×10^{-6}	-0.7249×10^{-6}
(−)		(−1.11)	(−1.00)
Monthly Debt	—	-0.1082×10^{-3}	-0.1094×10^{-3}
(+)		(−1.06)	(−1.00)
Years Employed	—	0.6349×10^{-2}	0.6387×10^{-2}
(−)		(0.89)	(0.94)
Self-Employed	—	−0.4374	−0.4531
(−)		(−2.99)[a]	(−3.08)[a]
Property Variable			
Age of House	—	0.3720×10^{-2}	0.3743×10^{-2}
(+)		(2.14)[a]	(2.15)[a]
Personal Variables			
Sex	—	—	0.1582
(−)			(0.87)
Marital Status	—	—	-0.9589×10^{-3}
(+)			(−0.007)
Race	—	—	0.3408
(+)			(1.45)[b]
Age	—	—	−0.1061
(+)			(−0.47)
Constant	−3.6036	−3.4832	−3.5150
Summary Statistics (−2.0 times *Log* Likelihood Ratio)	28.27	42.38	45.18
Number of Observations	3456	3456	3456

Note: The *t*-statistics are shown in parentheses under the coefficients; a one-tail test was employed.
[a] Significant at the 95 percent level.
[b] Significant at the 90 percent level.

impact on the lending decision with some caution.

REFERENCES

F. D. Nelson, "On A General Computer Algorithm for the Analysis of Models with Limited Dependent Variables," *Annals Econ. Soc. Measure.*, Fall 1976, *5*, 493–509.

New York State Banking Department, *Mortgage Financing and Housing Markets in New York State: A Preliminary Report*, New York 1977.

Differences in Unemployment Experience Between Blacks and Whites

By Charles L. Betsey*

Blacks currently represent about 12 percent of the working age population. During August 1977, black workers accounted for about 24 percent of the 6.9 million individuals who were unemployed. At that time, the black unemployment rate of 14.5 percent equalled its postwar high reached during 1975, and was about 2.4 times the 6.1 percent unemployment rate experienced by whites. Among black teenagers unemployment rates nationally remained at slightly more than 40 percent during the upturn, while the rate for white teenagers has declined to about 13 percent.

Historically, blacks have been about twice as likely as whites to be unemployed, almost irregardless of the level of overall economic activity. Among the various explanations given for the black-white unemployment rate differential are racial discrimination, differences in the level of educational attainment, differences in the quality of education, differences in the minimum wage acceptable for employment, and differences in job search patterns and job retention.

Using data for a low unemployment period (1969–70), my analysis indicates that among inner-city residents of New York City a racial disparity existed in unemployment experience and that while factors such as schooling, age, previous training, and other demographic characteristics mattered, they accounted for less than two-fifths of the variations in unemployment experience.

My results also indicate that the distinction between the number of spells of unemployment and the length of unemployment is an important one, and that the experience of blacks and whites differ substantially by either measure.

I. Previous Research

In a recent article, Nancy Barrett and Richard Morgenstern conclude from their analysis of unemployment rate differences by race and sex that the major unemployment problem faced by blacks and women is one of frequent job changes rather than chronic long-duration unemployment.

Other researchers, attempting to determine the cause for the relatively high overall rates of unemployment (with correspondingly high rates of inflation) prevailing in the United States in the past six to seven years, have also recognized the concentration of unemployment experience among blacks, women, and teenagers. Further, even for those groups which are relatively hard hit with unemployment, researchers have found the distinction between frequency of joblessness and duration of a spell of unemployment to be an important one, since the measured unemployment rate is not a unitary phenomenon.

It has been shown that the unemployment rate is the computational equivalent of the product of the number of spells of unemployment per person and the average duration of a spell. Thus the unemployment rate U/L is $N/L \cdot S \cdot D/52$, where N is the number of unemployed over the year, L is the labor force, S is the number of spells per person, and D is the average duration of a spell. Once the unemployment rate is decomposed into these components, it is a logical step to investigate whether differences in unemployment rates are caused more by variations in the average length of

*Congressional Budget Office (*CBO*). The views expressed are those of the author and do not necessarily represent the policies of the *CBO* or the views of other *CBO* staff members. This paper is a revised version of copyrighted material in my doctoral dissertation. I am grateful for the comments of Malcolm Cohen, Louis Ferman, William Neenan, and especially George E. Johnson on the original version. Finally, thanks to Norma Leake and Carolyn Levere who provided typing assistance.

spells of unemployment (duration), or by the number of spells of unemployment (turnover).

Based on findings that large differences exist among labor force groups in turnover rates, while differences in duration are relatively small by comparison, Robert E. Hall (1970) has concluded that turnover is likely the most significant factor in explaining high unemployment rates for blacks, women, and teenagers.

II. Duration of Unemployment

Since my interest is in explaining the determinants of unemployment rather than aggregate unemployment rates, I follow Hall (1970) in using total weeks unemployed as the measure of duration. Table 1 lists the results for variables significant for either group from my regressions of total weeks unemployed for black and white males aged 16 to 64. The reference group in each case is males 25 to 44 who are single, residents of East Harlem, and employed in nondurable manufacturing.

While the total duration of unemployment does not vary significantly with status as head of household among black males, white male heads of household experience significantly fewer weeks of unemployment than otherwise comparable whites. Among black males those who are separated experience unemployment of significantly longer duration than those who are married, divorced, or single. Other things equal, single white males experience the shortest spells of unemployment, with separated individuals experiencing the longest periods of joblessness. There are several possible factors which account for these findings relative to marital status. Among them is the possibility that separated males may tend to be of a lower socioeconomic background than divorced males and therefore more likely to have difficulty finding work.

The greatest duration of unemployment occurs for black males in the 16–19 age groups, with shorter durations for those age 45 and over. This finding is consistent with the hypothesis that younger people have more opportunity to search for jobs since they often lack pressing family obligations. It is also consistent with the idea that younger labor force members are likely to be new entrants and suffer from longer durations of unemployment as a consequence of inadequate information and unproductive methods of job search.

A differential pattern emerges for white and black teenagers, with white teenagers experiencing the fewest weeks of unemployment of any white subgroup, *ceteris paribus*, whereas black teenagers experience the most weeks unemployed. Among other age groups duration of employment for whites is generally lower than

TABLE 1—WEEKS OF UNEMPLOYMENT

Variable	Black Males	White Males
Head	.420	−.199
	(.440)	(.080)[a]
Married	.057	.120
	(.384)	(.069)[c]
Separated	1.316	.519
	(.377)[a]	(.118)[a]
Age 16–19	.517	−.257
	(.766)	(.143)[a]
Weeks Worked	.033	−.008
	(.019)[c]	(.004)[b]
Wage Rate	−.136	−.015
	(.069)[b]	(.010)
Finance	−2.064	.419
	(2.691)	(.252)[c]
Services	−1.139	.444
	(1.717)	(.221)[b]
Central Harlem	−1.329	−.061
	(.462)[a]	(.117)
South Bronx	−1.265	.297
	(.487)[a]	(.100)[a]
Brooklyn I	−1.327	−.028
	(.552)[b]	(.069)
Brooklyn II	−1.261	−.075
	(.444)[a]	(.110)
Brooklyn III	−1.069	−.019
	(.538)[b]	(.063)
Spells	2.229	.148
	(.301)[a]	(.057)[a]
Constant	−.388	.415
$\bar{R}^2$	.097	.087
N	691	521
F	3.43[a]	2.63[a]

[a] Significant at the 1 percent level.
[b] Significant at the 5 percent level.
[c] Significant at the 10 percent level.

that of blacks, particularly in the 25–44 year age range.

Number of weeks worked in the previous year is a significant variable in predicting duration of unemployment for both white and black males. The results obtained, however, indicate that black males are most subject to unstable employment, since the number of weeks worked in the previous year is positively correlated with the number of weeks of unemployment black males can expect in the current year. For white males just the opposite is true, greater stability in the previous year pays off in fewer weeks of unemployment in the current year. The differential effect of the weeks worked variable lends indirect support to the hypothesis that black unemployment is partially a result of unstable and unsatisfactory jobs, while for white males unemployment is of a more normal frictional nature.

When controlling for other factors it appears that both groups of males experience fewer weeks of unemployment as their hourly wage rates increase. Among black males the duration of unemployment is substantially reduced at higher wage rates, indicating that the longest periods of joblessness occur for black males who are low wage workers. While black unemployment duration did not vary significantly with industry attachment or unionization, expected duration of unemployment among white males was greatest for those employed in services and finance, insurance, and real estate.

Area of residence was a significant factor in predicting differences in weeks unemployed among black males; residents of all areas had significantly lower durations than East Harlem residents. Among white male labor force participants, only those who resided in the South Bronx showed significant differences in weeks unemployed; about 0.3 weeks longer than the duration of comparable East Harlem residents.

The number of completed spells of unemployment in the previous year exerts a significant influence upon current duration of unemployment. For black males, each additional spell of unemployment increases the duration of unemployment by about eleven days. Similar increases in the number of spells affect white males' duration barely one-fifth as much: an additional previous spell of unemployment results in current duration for whites being longer by one day.

The finding relative to the importance of the spells of unemployment variable is significant for several reasons. First, my results clearly show that previous spells of unemployment are a significant predictor of weeks of unemployment. Hall (1970) predicted weeks of unemployment as a function of several demographic and background variables, but omitted spells of unemployment. Given my findings, his analysis seems to have omitted the most important variable, though, in general, his findings are similar to my own. Secondly, the differential effect of spells of unemployment by race indicates that black males' more frequent spells of joblessness are associated with their experiencing longer periods of unemployment in the future than otherwise similar workers.

III. Spells of Unemployment

In an attempt to understand why black and white turnover rates differ I regressed spells of unemployment on the several demographic variables used earlier. Table 2 presents the results for those factors which were significant for either group. The ability to explain spells of unemployment is considerably better for black males than for whites.

Before discussing these results, it may be instructive to draw a distinction which is rarely made clear in the literature. In Hall (1972), and the work of Barrett and Morgenstern and others, the analysis of spells of unemployment is often discussed relative to the reasons some workers, blacks in particular, change jobs more often than others. The analysis of turnover, in the sense of spells of unemployment, should not be confused with turnover in the sense of change of jobs. Individuals sometimes change jobs without becoming unemployed

TABLE 2—SPELLS OF UNEMPLOYMENT

Variable	Black Males	White Males
Age 16–19	−.400	−.215
	(.098)[a]	(.113)[c]
Suburb	.103	.096
	(.056)[c]	(.071)
Union	3.80	4.19
	(.840)[a]	(.569)[a]
Agriculture	1.60	1.27
	(.521)[a]	(.440)[a]
Construction	.084	−.569
	(.079)	(.114)[a]
Durable	−.027	−.206
	(.068)	(.079)[a]
Transport	−.133	.722
	(.077)	(.125)[a]
Trade	.862	.958
	(.193)[a]	(.137)[a]
Finance	1.56	1.24
	(.324)[a]	(.192)[a]
Services	.956	1.16
	(.219)[a]	(.167)[a]
Government job	.375	.116
	(.104)[a]	(.069)[c]
Bronx	−.105	.103
	(.063)[c]	(.079)
Brooklyn I	−.127	−.033
	(.071)[c]	(.055)
Brooklyn II	−.112	.040
	(.057)[b]	(.087)
Constant	−1.627	−1.438
$\bar{R}^2$	.332	.245
N	691	521
F	12.47[a]	6.67[a]

[a] Significant at the 1 percent level.
[b] Significant at the 5 percent level.
[c] Significant at the 10 percent level.

and may become unemployed though not changing jobs (for example, through layoffs). To the extent that this occurs, the interpretation of job change as unemployment or vice versa is misleading. Besides, there are factors which may affect the probability of an individual experiencing a spell of unemployment although they may be unimportant in determining whether or not an individual changes jobs.

This is not to argue that for some or even most workers changing jobs does not involve a high likelihood of unemployment. That is an empirical question which is beyond the scope of my inquiry. Yet, it is important to distinguish between job changes and unemployment spells. It is interesting to note that this distinction can lead to somewhat different interpretations regarding empirical findings. For example, in discussing the findings that income is negatively related to spells of unemployment some authors have argued that low income workers occupy undesirable jobs (not at all unlikely), and that as a consequence they change jobs more often in search of better job situations. Though this may be true, the results afford no direct support of this assertion. The finding is also consistent with the hypothesis that higher wage and/or higher income individuals are less likely to experience a spell of unemployment when changing jobs because they have more skills, more reliable job information, and more opportunity to search while presently employed.

My results show that teenagers of both races experience significantly fewer periods of joblessness than other age groups, other things equal. This finding runs contrary to expectations and also seems inconsistent with previous studies on the behavior of teenage labor force participants. Results reported by George Perry, and Barrett and Morgenstern, indicate that teenagers move into unemployment much more frequently than other age groups and generally account for a disproportionate amount of total unemployment.

Hall (1970) obtained results similar to mine when estimating probabilities of becoming unemployed using the Survey of Economic Opportunity. In explaining his results for teenagers, Hall observes that information on unemployment experience obtained in the major household surveys is often supplied by a household member other than the teenager himself, and therefore the frequency and duration of teenage unemployment is probably understated in these surveys.

My data suffer a similar fault since they were obtained in the same manner. Yet there is another plausible explanation for this apparent anomaly. Hall (1970), as I do, uses multiple regression techniques to hold constant the various demographic and locational variables in the analysis. It would ap-

pear that controlling for such influences as marital status, headship, and wage rates substantially reduces the expected number of spells of unemployment for male teenagers.

The control for wage rate seems particularly important in this regard. My own previous work has shown that wage rates vary substantially depending upon the occupation and industry of employment. Teenagers, due to age, education, and years of work experience, are overrepresented in the lower skilled occupations, which tend to have high turnover rates. Once these factors are controlled for, teenage turnover rates are lower than those of prime-age males.

Black inner-city residents who commute to suburban areas are more likely to have spells of unemployment than those who work in the poverty area or some other part of the city. For white workers the opposite tendency exists; those who work outside the city boundaries experience a somewhat lower incidence of joblessness. While the size of the effect is minimal in both cases (suburban employment raises the number of spells by about one-tenth for blacks and reduces them by the same amount for whites), the significantly positive result for blacks is interesting.

Contrary to my expectation that unionism enhances job security and results in a reduced incidence of unemployment, it was found that higher levels of unionization are associated with increased numbers of spells of unemployment among white and black workers alike. At the same time, the increase in unemployment incidence accompanying unionism is greater for whites than blacks.

The results with respect to unionization are distinct from the unemployment incidence associated with industrial attachment (to the extent that the two influences are distinguishable). For both black and white males spells of unemployment are most frequent, other things equal, in agriculture, fisheries, and forestry-based employment, while the lowest incidence of unemployment for black males occurs in transportation, communications, and public utilities. Public sector employment for residents of New York's inner-city neighborhoods is found to be less stable than employment in manufacturing.

Finally, while area of residence is not a significant determinant of unemployment for white males, black male residents of the Bronx and some areas of Brooklyn report significantly fewer periods of joblessness than Harlem residents.

IV. Summary

My findings indicate that the determinants of unemployment differ significantly among black and white male labor force participants, whether our focus is the frequency of unemployment (spells) or the length of unemployment (duration). Other things equal, blacks' unemployment duration increases considerably more with an earlier spell of unemployment than that of comparable whites. Among blacks, each spell of unemployment results in about two additional weeks of future joblessness; for whites, on average, each occurrence results in a day's future job loss.

My finding of a differential effect of spells of unemployment by race is of considerable interest from a policy standpoint. If my interpretation is correct, black males are affected by a series of factors, race and previous spells of unemployment among them, which result in much longer periods of unemployment than for otherwise similar workers.

Increases in the number of spells of unemployment seem to increase the length of future jobless periods differentially for blacks and whites. If this is indeed the case, then much might be gained if spells of unemployment can be reduced, and black workers would stand to gain more than whites in terms of potential weeks of employment.

REFERENCES

W. H. L. Anderson, "Trickling Down: The Relationship Between Economic Growth and the Extent of Poverty Among Amer-

ican Families," *Quart. J. Econ.*, Nov. 1964, *78*, 511–24.

N. S. Barrett and R. D. Morgenstern, "Why Do Blacks and Women Have High Unemployment Rates?," *J. Hum. Resources*, Fall 1974, *9*, 452–64.

C. L. Betsey, "Work Experience and Earnings of Inner-City Blacks and Whites," unpublished doctoral dissertation, Univ. Michigan 1976.

H. J. Gilman, "Economic Discrimination and Unemployment," *Amer. Econ. Rev.*, Dec. 1965, *55*, 1077–95.

R. E. Hall, "Why Is The Unemployment Rate So High At Full Employment?," *Brookings Papers*, Washington 1970, *3*, 369–402.

———, "Turnover in the Labor Force," *Brookings Papers*, Washington 1972, *3*, 709–56.

B. Harrison, "Education and Underemployment in the Urban Ghetto," *Amer. Econ. Rev.*, Dec. 1972, *62*, 796–812.

G. Iden, "Business Conditions, Demography, and the Teenage Unemployment Problem," paper presented at the Southern Economic Association Meetings, New Orleans, Nov. 1977.

G. L. Perry, "Unemployment Flows in the U.S. Labor Market," *Brookings Papers*, Washington 1972, *2*, 245–78.

R. E. Smith and C. C. Holt, "A Job-Search Turnover Analysis of the Black-White Unemployment Ratio," work. paper no. 350-426, Urban Instit. 1971.

H. M. Wachtel, and C. Betsey, "Employment at Low Wages," *Rev. Econ. Statist.*, May 1972, *54*, 121–29.

U.S. Congress, Congressional Budget Office, *The Unemployment of Nonwhite Americans: The Effects of Alternative Policies*, Washington 1976.

U.S. Bureau of the Census, *Employment Characteristics of Selected Low-Income Areas*, PHC(3), Washington 1973.

U.S. Bureau of Labor Statistics, "The Employment Situation: August 1977," *News*, Washington, Sept. 2, 1977.

DISCUSSION

KARL D. GREGORY, Oakland University: In an excellent analysis of racial differences as one approach to viewing unemployment, Charles Betsey focuses upon males, aged 16–64, in New York's inner city. He employs data for 1969–70, a period of relatively low unemployment. Following tradition, he factors the unemployment rate into three components, including the number of spells of unemployment and the average duration of each spell. Through multiple regression analysis of each of these two components—number of spells and duration—for males, individually by race, Betsey finds major differences among black and white males in their unemployment experiences. His models differ in part from those of earlier studies and cast new light on racial disparities.

Among the conclusions I found most interesting are the following: for black males, several characteristics are associated positively with a longer duration of unemployment. Among these characteristics are being a family head, separated, age 16–19, or having several spells of unemployment in the previous year.

The crisis in black teenage unemployment finds confirmation in this study. Black teenage males suffer greater duration of unemployment than any other group. In contrast, white teenage males experience the shortest duration of any white subgroup.

For all age groups, the duration of unemployment is higher for blacks than for whites. Presumably, this added burden for blacks would have been longer were the data corrected to adjust for periods of joblessness caused by worker discouragement.

In contrast with conclusions from previous studies by Robert Hall, and by Nancy Barrett and Richard Morgenstern, Betsey finds striking differences among black and white males in both duration and spells of unemployment. He also found union membership is strongly and positively associated with spells of unemployment for both black and white males.

One finding has especially great implications for public policy. Spells of unemployment in one year are associated with an increase in the duration of unemployment of blacks in the following year by two weeks per spell, and for whites, by one day per spell, a tenth as much as for blacks. Similarly, if spells in one decade are related to spells and added duration of unemployment in future decades, as seems plausible, the current sustained high level of unemployment, particularly for blacks, would have great continuing consequences for decades to come. Moreover, the social benefits from lowering unemployment now would be greater than current calculations comprehend. Another implication for public policy is that, *ceteris paribus*, any given reduction in unemployment will reduce transfer payments associated with unemployment by a greater extent, the larger the proportion of the added jobs occupied by blacks. One must however be mindful that this is a study of only one city. The results may not be completely generalizable.

Several other, largely technical, comments appear warranted. The unexplained variance in the regression model for the duration of unemployment may dominate the explained variance. The $\bar{R}^2$ was less than 10 percent for both black and white males. Also, there is probably colinearity among some of the variables. The severe space limitations undoubtedly precluded Betsey's discussion of such fine points.

One factor which was not discussed could conceivably bias greatly the analysis by understating the already large racial disparities. The discouraged worker effect may conceal some periods of unemployment to the extent that the discouraged worker is correct in his expectation of a dearth of jobs for himself.

The interactions over time between spells of unemployment and duration of unemployment suggest that a single equation approach may not capture adequately key relationships. A system of simul-

taneous equations would appear to be a superior approach.

The title is quite misleading, for racial differences in unemployment experiences suggests an inquiry extending beyond a mere examination of spells and the duration of unemployment. This might be remedied partly in further research by extending the analysis to a separate examination of the various reasons for unemployment, so that variations in the force of the explanatory variables may be observed for job losers, job leavers, reentrants, and new entrants. Such an extension could be helpful for selective manpower policies.

A Partial Survey of Recent Research on The Labor Supply of Women

By JAMES J. HECKMAN*

This paper is a progress report on some recent research on the labor supply of married women. In previous work on this topic two conceptually distinct interpretations of the coefficients of regressions of labor force participation status (measured as a rate for a group or by a dummy variable for an individual) on wage rates and unearned income have been presented. The first interpretation, which stems from Jacob Mincer's pioneering work and which is pursued in later work by Marvin Kosters, Glen Cain, and Orley Ashenfelter and the author, interprets estimated wage and income coefficients as estimates of substitution and income effects. The later work in this tradition interprets the income and substitution effects within the Hicks-Slutsky framework. The operating assumption in this literature is that the estimated wage and income coefficients from an analysis of participation admit the same theoretical interpretation as wage and income coefficients obtained from hours of work regressions. The economic model that yields such an interpretation is a life cycle model of labor supply.

A second interpretation, that has been the guiding principle of much recent work on labor supply, stresses the point that labor force participation at a point in time is a discrete decision. A woman either participates or does not participate, and estimates of participation equations are conceptually distinct from estimates of hours of work functions. Participation regressions describe "corner phenomena" and do not estimate "interior solution" Hicks-Slutsky income and substitution effects, although they certainly estimate parameters of the utility function of consumers. The modern statement of the second view has been developed in papers by Yoram Ben-Porath and H. Gregg Lewis which were important stimuli to later work on female labor supply by Reuben Gronau, the author (1974), J. Cogan, and Giora Hanoch.

While the development of the second approach is still underway, it is important to note that at the present time, analysts operating within the second tradition have ignored the focus on the *life cycle* that underlies the first tradition. Analysts operating within the second tradition assume that their one-period models of labor supply apply to the data used to test their models—typically hours of work or weeks of work in a survey year—and ignore the point that most consumers have ample opportunity to substitute time and goods over the life cycle, and to invest in human capital.

This paper presents some of the principle results of recent research on life cycle labor force participation that attempts to merge these two traditions. In order to place this work in a suitable perspective, it is necessary to review some of the previous literature. In this review, certain implicit assumptions in previous work are spelled out for what is believed to be the first time. Given space limitations, only informal arguments are presented. A more complete analysis is available in a companion paper (see the author, 1977a).

*University of Chicago. This work was partially supported by a U.S. Department of Labor grant to the National Bureau of Economic Research. I have greatly benefited from discussions with T. MaCurdy and J. Mincer. Any remaining errors are mine.

I. Mincer's Model and Subsequent Developments

The basic Mincer model is presented in a few lines of his seminal paper:

> In a broad view, the quantity of labor supplied to the market by a wife is the fraction of her married life during which she participates in the labor force. Abstracting from the temporal distribution of labor force activities over a woman's life, this fraction could be translated into a probability of being in the labor force in a given period of time for an individual, hence into a labor force rate for a large group of women. [p. 68]

Mincer goes on to develop a model of the timing of labor force activity over the life cycle. The implicit model that underlies the econometric models of Mincer, Cain, and Ashenfelter and the author is based on a lifetime utility model of consumer decision making. Since this model has never been explicitly developed, it seems useful to do so here to emphasize the principles involved.

Lifetime utility is a well-behaved function of lifetime consumption of goods X and leisure L. The maximum amount of leisure that can be consumed is T. Let ϵ be an unobserved "taste" or "household production" variable assumed independent of other variables. Lifetime utility is $U = U(X, L, \epsilon)$. The consumer faces a permanent real wage W net of all money costs of work. The price of goods is unity. Assume a zero interest rate. Given resources A, the consumer works sometime in his life if

$$M(A,T,\epsilon) = \frac{U_2(A,T,\epsilon)}{U_1(A,T,\epsilon)} \leq W \tag{1}$$

where U_j is the partial of U with respect to argument j (assumed here to be positive for all variables). The term on the left of the inequality is the marginal rate of substitution between goods and leisure at full leisure, i.e., the lifetime reservation wage.

If the consumer works, labor supply is determined by the solution to the equations

$$M(X,L,\epsilon) = \frac{U_2(X,L,\epsilon)}{U_1(X,L,\epsilon)} = W \tag{2a}$$

$$X + WL = WT + A \tag{2b}$$

The equations can be solved for the fraction of total time worked h as $h = (T - L)/T = h(W, A, \epsilon)$. Denoting h_j as the partial of h with respect to its jth argument, h_1 is the ordinary (uncompensated) Slutsky effect of permanent wage change on labor supply and h_2 is the income effect.

By construction, the simple theory is silent on the timing of labor supply over the life cycle. Ben-Porath and others interpret Mincer's paper as enriching the simple theory by assuming that, aside from "transitory" factors (for example, factors like children, transitory income variation, and the like), the timing of participation over the life cycle is "random." The probability of finding a randomly selected consumer of a group of consumers with permanent wage W and assets A in the labor force at any point in time is the population mean of h. *Assuming that all consumers work sometime in their life cycle*, regressions of a participation measure on variables W and A estimate the mean values of h_1 and h_2. These estimated coefficients may be interpreted as estimates of Hicks-Slutsky substitution and income effects.

It is not necessary to assume that observationally identical individuals have identical participation probabilities. Depending on the distribution of the unobserved variable ϵ in the population, there may be considerable dispersion in tastes, or unobserved permanent components of wage rates, leading to a distribution in population probabilities. The only crucial assumptions are 1) that the participation rate measures the fraction of lifetime labor supplied to the market, and 2) that the fraction of lifetime labor supplied to the market is generated by an *interior solution* to the lifetime utility-maximization problem. Assumption 1 is quite weak and is satisfied in a wide variety of situations while assumption 2 is much more demanding. The only requirement for assumption 1 to be valid is that the economic environment is stationary or, if it is not, that it be changing in a way that is known or can be determined.[1] In particular one must be able to control for cyclical and cohort effects.

[1]Thus Ben Porath's dichotomy between homogeneity and heterogeneity in his analysis of hypothesis

Ben-Porath has recently noted that many women do not participate in the labor force over substantial time intervals (ten years in his data). Later work by the author and Willis suggests that the proportion of married women who never participate is substantial in short intervals of data. However, a recent note by Mincer and Haim Ofek suggests that the fraction of married women who never participate (after leaving school) is only 4–5 percent when lifetime labor supply is measured (however, in their own data, 25 percent of women do not work after they are married). Given that *some* women never work, what is the implication of this finding for the interpretation of participation regressions?

The answer to this question is simple, although it is not in the literature. The probability that a randomly selected member of a group with permanent wage W and asset position A *ever* works is:

$$Pr(W \geq M(A, T, \epsilon)) = P(W, A)$$

where the probability is computed from the population distribution of unobservable variables. The expected probability that a randomly selected individual with wage W and assets A who ever works will be found working at a randomly selected point in time is $E_c(h) = E[h(W, A, \epsilon) \mid W \geq M(A, T, \epsilon)]$ where $E_c(h)$ is the fraction of lifetime hours that are worked by an ever-working individual. Recent work in labor supply demonstrates that $\partial E_c(h)/\partial W \neq h_1$ because the left-hand side term includes the effect of movement across taste distributions as the wage is raised, while the right-hand side term is the Slutsky effect holding tastes constant. The left-hand term includes the effect of an increase in wages on the mean unobservable "taste for work" for those who work while the right-hand side term does not.

The probability that a randomly selected individual from a population of people with permanent wage W and asset position A works at a point in time is

I and hypothesis II in his paper is quite misleading. The real issue is "interior solution vs. corner solution" and not homogeneity vs. heterogeneity.

$$P(W, A)\, E_c(h) \tag{3}$$

If everyone works sometime in the life cycle, as Mincer maintains, $P(W, A) = 1$ for all values of W and A and $E_c(h) = E(h)$. Regression estimates of participation status on wage rates and assets estimate h_1 and h_2, the Hicks-Slutsky terms.

If the population is dichotomized into full-time workers and full-time nonworkers, $E_c(h) = 1$, and participation regressions estimate the parameters of $P(W, A)$, i.e., the parameters of M, the reservation wage function at full leisure, and the parameters of the distribution ϵ. (This is Ben-Porath's hypothesis II.)

In the general case, cross-section participation equations estimate the first partials of expression (3) which are *not* h_1 and h_2, the Hicks-Slutsky terms, and which are *not* P_1 and P_2, parameters derived from the marginal rate of substitution function M and the parameters of the distribution of ϵ. Estimates of the parameters of expression (3) yield the parameters of the aggregate labor supply function for women in terms of permanent wage rates and assets, since this expression combines the effect of permanent wage and asset changes on the entry of women into the market and on hours supplied by previously working women. The aggregate supply function can be used to predict the probability that a randomly selected woman with wage W and assets A will be in the market at a point in time, and it can be used to predict the fraction of her lifetime that a randomly selected woman with these traits will work. Thus the agreement in the estimates of the permanent wage elasticities for these two measures of labor supply reported by Cain (p. 100) is reassuring.

The simple theory just discussed is not entirely satisfactory for three reasons: it rests on certain implicit assumptions which when brought to light make it somewhat unpalatable; it does not provide a natural framework for integrating different dimensions of labor supply activity in a unified model; and it is easy to misuse it as a guide to dynamic models of labor supply.

The assumption about the source of ran-

domness that generates the timing of the work decision is critical in interpreting estimates of the response of participation to wage rates and assets as Hicks-Slutsky effects, even if everyone works. One plausible source of variation in the timing of labor supply over the life cycle is differences in transitory components of wage rates known to the consumer but unknown to the observing economist.[2] This source is mentioned in Mincer's work. If consumers are rational, and if wages (or shadow costs) vary over the life cycle, even if all work sometime in their lifetime, regressions of participation status on permanent wages and assets *do not* estimate income and substitution effects.

To see this, break life up into T periods with wages $W_1, \ldots, W_T$. Denote the largest wage by W^1 and the smallest by W^T. The permanent wage can be defined as the average wage over the life cycle, as is customary. The decision to ever participate is made by comparing the lifetime reservation wage to the highest wage W^1 and *not* the permanent wage as in inequality (1). The number of periods a consumer works is not determined from equations (2a) and (2b). In fact, a consumer works k periods if

$$\text{(4a)}\quad M\left(A + \sum_{i=1}^{k-1} W^i, T - k + 1, \epsilon\right) \leq W^k$$

$$\text{(4b)}\quad M\left(A + \sum_{i=1}^{k} W^i, T - k, \epsilon\right) > W^{k+1}$$

While the permanent wage is a determinant of these inequalities, it does not play the same role in them as in equations (2a) and (2b). The marginal wage relevant to the participation decision, and the decision to work k periods, is not the average wage. The average wage for the periods in which the consumer works, which is relevant for evaluating "income effects," exceeds the marginal wage relevant for evaluating "substitution effects" in this model.

Given the joint distribution of wages, and ϵ (assumed independent of wages), inequalities (4a) and (4b) can be used to derive the probability that the consumer works $t = 0, \ldots, T$ periods as a function of permanent wage W, and permanent asset position A; $P_t = P_t(W, A)$, $\sum_{t=0}^{T} P_t = 1$. The partial derivative of each P_t function with respect to W does *not* directly estimate a Hicks-Slutsky income and substitution effect. However, utilizing a direct extension of the statistical model of Richard Rosett and Forrest Nelson, estimates of the P_t functions can be used to piece together segments of the lifetime utility function at $T + 1$ intervals so that rough estimates of the consumer's indifference system may be obtained. These estimates can be used to estimate Hicks-Slutsky income and substitution effects.

The probability of finding a randomly selected member of the population with permanent wage W and asset position A in the labor force in any period, which is also the expected proportion of life that an individual will work, is

$$\text{(5)}\qquad \sum_{t=0}^{T} \frac{t\, P_t(W, A)}{T}$$

The partial derivatives of this function are what is estimated in cross-section regressions of participation status on permanent wage rates and assets, and these partials are clearly *not* the usual income and substitution effects.

What type of "randomness" will lead to the simple model employed in the previous literature? One possibility is a random utility model (see Daniel McFadden) in which the consumer's work decision in each period is made by tossing a coin with probability $h (0 < h < 1)$ of working (assuming T is "large"). But transitory wage or cost variation will not lead to this model if the consumer is rational in the neoclassical sense of that term.

Estimates of equation (5) from the cross section cannot be used to estimate the distribution of participation sequences in the population or the distribution of the number of periods ever worked over the life cycle (given by the component parts of equation (5)) unless the random utility model is

[2] A similar analysis can be made if there are unobserved deviations in household production or tastes over the life cycle.

assumed, successive tosses of the coin are independent, and individuals are identical (the distribution of ϵ is degenerate). In the general case, one cannot use the cross-section mean to estimate the probability of any sequence of participation decisions over the life cycle if there is any unobserved heterogeneity in the population (see the author and Willis).

II. Dynamic Models of Labor Supply: Some Preliminary Results

In this section, I present the most elementary version of a class of dynamic models of labor supply that have recently been estimated by Thomas MaCurdy and myself. This model can be used to interpret the interrelationship among the various dimensions of labor supply analyzed in the literature and can also shed some revealing light on certain aberrant empirical findings in the "new" labor supply literature that rigorously analyzes one-period models of labor supply and applies them to the data at hand—typically a survey year or survey week.

The consumer's utility at age t is a strictly concave, twice differentiable nonsatiable function $U(t) = G(X(t)) + J(L(t))$, $t = 0, \ldots, T$, where $X(t)$ is the consumption of goods and $L(t)$ is the consumption of leisure at age t. The maximum value of leisure is 1. The real wage is $W(t)$ and the real price of goods is unity. In this simple model, human capital accumulation is ignored, although it is not ignored in the more sophisticated models that we have estimated. Assuming no uncertainty, a parametric interest rate r and rate of time preference ρ, the consumer maximizes

$$\sum_{t=0}^{T} U(t)(1+\rho)^{-t}$$

subject to the budget constraint

$$A = \sum_{t=0}^{T} (1+r)^{-t} \cdot [W(t)[1 - L(t)] - X(t)] \tag{6}$$

where A is initial wealth. This model is essentially a finite horizon Ramsey model.

For ages at which the consumer works, the demand functions for goods and leisure are

$$X(t) = X\left(\lambda\left(\frac{1+\rho}{1+r}\right)^t\right) = [G']^{-1}\left(\lambda\left(\frac{1+\rho}{1+r}\right)^t\right)$$

$$L(t) = L\left(\lambda W(t)\left(\frac{1+\rho}{1+r}\right)^t\right) = [J']^{-1}\left(\lambda W(t)\left(\frac{1+\rho}{1+r}\right)^t\right)$$

where $X' < 0$, $L' < 0$, and $\lambda\ (> 0)$ is the marginal utility of wealth (the LaGrange multiplier associated with the budget constraint). Goods and leisure are assumed to be normal.

If the consumer does not work at age t, $L(t) = 1$, and

$$\frac{J'(1)}{\lambda}\left(\frac{1+\rho}{1+r}\right)^t > W(t) \tag{7}$$

The term on the left-hand side of the inequality is the shadow price of time at age *t evaluated at the equilibrium position*. The term on the right-hand side is the market wage. If the equilibrium reservation wage exceeds the market wage, the consumer does not work in period t.

The value of λ is determined jointly with the other decision variables by substituting the demand functions into resource constraint (6). Note that λ is a function of all the parameters of the model including the vector of lifetime wage rates.[3] Lifetime labor supply may be defined as the sum of the number of periods worked or the number of hours worked—two conceptually distinct measures which are related to the underlying set of parameters in different ways.

I briefly consider some of the implications of the model. They are spelled out in greater length in a companion paper (see

[3]For any monotonic transformation of the utility function $\Sigma\, U(t)\,(1+\rho)^t$, λ defined in this way is invariant.

the author, 1977a). First consider the participation decision. Recent work on female labor supply utilizes inequality (7) to characterize the participation decision. The key assumption in that literature is that the market wage is independent of the reservation wage. In a one-period model of labor supply, this assumption is appropriate. It is also an appropriate assumption for determining the probability that a consumer will *ever* work in a life cycle model. However it is an inappropriate assumption for characterizing the participation decision in a given period if the consumer works in other periods, and if the wage in the given period is correlated with the wage in other periods, as G. Sedlacek has found is the case for women. Under the maintained assumptions, if consumers work in other periods, higher current wages are associated with lower values of λ and higher values of the reservation wage. It is thus possible that wage rates are inversely correlated with labor force participation at a point in time.

This point has bearing on some recent findings by R. Olsen and James Smith. They find that for certain groups of women, *lower* wage women are the ones more likely to participate. This finding can be explained in a one-period model by appealing to a large positive correlation between tastes for work and unobservable determinants of wages. The life cycle model can explain this fact by appealing to intertemporal correlation in wage rates and the plausible assumption that women who work today may also work some time in the future. It is significant that the "perverse" association between wage rates and participation status is found in demographic groups with the greatest volume of lifetime labor supply—such as married black women. It is in such groups that income effects are likely to be the largest.

This point also implies that the estimated value of the reservation wage or value of home time obtained through the procedures of Gronau and the author (1974) understates the true value of time at home for women with higher market wages, and overstates it for women with lower market wages, since a higher permanent wage leads to a higher value of time. The Gronau-Heckman procedures hold the market wage component of the reservation wage function fixed at the sample mean and thus understate the true *equilibrium* reservation wage for women with higher wages.

The crucial identifying restriction utilized in their work—that certain variables that determine market wages are excluded from the reservation wage equation—does not hold in a life cycle model in which some women work in periods other than the one under investigation. Only market wage variables that do not appear in other periods can serve as identifying variables and this is a limited set of variables.

Second, the probability that a randomly selected consumer will be in the labor force in a given period can be modeled as a discrete choice problem. The consumer can be in 2^{T+1} possible life cycle states corresponding to each possible participation sequence. For each sequence, there is a lifetime utility and the consumer is assumed to pick the sequence with the highest utility.[4] From the probability of different participation patterns one can derive the probability of participation at a point in time. The effect of permanent wages on participation is not a Hicks-Slutsky effect nor is it a corner solution participation effect. There are wealth effects of interpersonal wage differences that arise in evaluating the relative utility of different participation sequences.

Third, a frontal assault on the estimation of the 2^{T+1} choice discrete choice model leads to very difficult data requirements and intractable estimation problems. One can avoid these complications by following a suggestion of MaCurdy and noting that for a given consumer, λ stays fixed over the life cycle. Given λ, one can characterize consumption, labor supply, and participation decisions in terms of estimable func-

[4]This involves fixing hours of work at zero in certain periods while letting them be freely chosen in the remaining periods. Note that some sequences may be mathematically impossible to realize for certain values of the unobservable variables and for certain functional forms for preferences.

tions and current values of the variables. Thus, a simple estimation method consists of treating λ as a *fixed effect* for a consumer, and estimating λ and the parameters of the "λ constant" demand functions written above using longitudinal data. In the case of the life cycle participation decision, this leads to a multivariate probit model with fixed effects (see the author, 1977b). As I have noted elsewhere, given λ, the parameters of $[G']^{-1}$ and $[J']^{-1}$ and the budget constraint, one can determine all the parameters of the life cycle labor supply and consumption functions. Models of this sort have been estimated by MaCurdy and myself. By focusing on the parameters of the preference function, we generate the *entire* set of parameters of the life cycle labor supply functions.

An important feature of the λ constant demand system is that it permits one to avoid the inevitably *ad hoc* definitions of permanent and transitory wage and asset change, and it permits estimation of the effect of current wages as direct determinants of behavior (holding λ fixed) and as determinants of λ.

One can also parameterize λ as a function of observable variables. Then the proper specification of the labor supply equations requires current and future values of variables as determinants of current labor supply activity. In other work (see the author, 1977a), I demonstrate that future variables, such as children, unemployment of the husband, and the like, affect current behavior. In particular, estimates of the hours of work equation that omit future wages *understate* the effect of current change in wage rates on current labor supply.

REFERENCES

O. Ashenfelter and J. Heckman, "The Estimation of Income and Substitution Effects in a Model of Family Labor Supply," *Econometrica*, Jan. 1974, *42*, 73–86.

Y. Ben-Porath, "Labor Force Participation Rates and the Supply of Labor," *J. Polit. Econ.*, May/June 1973, *81*, 697–704.

Glenn Cain, *The Labor Force Participation of Married Women*, Chicago 1966.

J. Cogan, "Labor Supply with Time and Money Costs of Participation," The Rand Corp., Aug. 1976.

R. Gronau, "The Effect of Children on the Housewife's Value of Time," *J. Polit. Econ.*, Mar./Apr. 1973, *80*, S30–S47.

G. Hanoch, "Hours and Weeks in the Theory of Labor Supply," The Rand Corp., R-1787, July 1976.

J. Heckman, "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

———, (1977a) "Dynamic Models of Female Labor Supply," unpublished paper, Univ. Chicago, Apr. 1977.

———, (1977b) "Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence," unpublished paper, Univ. Chicago, July 1977.

——— **and R. Willis,** "A Beta Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *J. Polit. Econ.*, Feb. 1977, *85*, 27–58.

M. Kosters, "Effects of an Income Tax on Labor Supply," in Arnold C. Harberger and Martin J. Bailey, eds., *The Taxation of Income from Capital*, Washington 1969.

H. G. Lewis, "On Income and Substitution Effects in Labor Force Participation," unpublished paper, Univ. Chicago 1967.

T. MaCurdy, "Labor Supply Decisions over The Life Cycle," unpublished paper, Univ. Chicago 1977.

D. McFadden, "Economic Applications of Psychological Choice Models," unpublished paper, Univ. California-Berkeley 1975.

J. Mincer, "Labor Force Participation of Married Women: A Study of Labor Supply," in Harold G. Lewis, ed., *Aspects of Labor Economics*, Princeton 1962.

——— **and H. Ofek,** "The Distribution of Lifetime Labor Force Participation of Married Women," unpublished paper,

Columbia Univ., Aug. 1977.

R. Olsen, "An Econometric Model of Family Labor Supply," unpublished doctoral dissertation, Univ. Chicago, June 1977.

R. Rosett and F. Nelson, "Estimation of the Two-Limit Probit Regression Model," *Econometrica*, Jan. 1975, *43*, 141–46.

G. Sedlacek, "Dynamic Models of Female Wage Growth," unpublished paper, Univ. Chicago 1977.

J. Smith, "The Convergence to Racial Equality in Women's Wages," The Rand Corp., Sept. 1977.

Fertility and Child Mortality over the Life Cycle: Aggregate and Individual Evidence

By T. PAUL SCHULTZ*

Mortality has been markedly reduced in the poorer countries of the world in the last three to four decades; existing data typically indicate life expectancy increasing about one-half year per calendar year during this period (see G. J. Stolnitz). This decline of mortality has contributed to an increase in intrinsic rates of population growth of almost 1 percent per year. Further, since the most dramatic changes are observed for infants and children, the age composition of these populations has changed, contributing today to a transitory increase in the rate of population growth of another one-half of 1 percent per year.

In this unprecedented period of rapid population growth, a natural question to ask is whether fertility responds to the decline in mortality, and if so, then by how much and how fast, and whether this tendency toward demographic equilibrium dampening the rate of population increase across countries is also evident across economic classes within countries. There are some indications that the reduction in mortality may have been concentrated in lower income groups and though this would be appropriately construed as an egalitarian development (see Simon Kuznets), it also raises the possibility that the rate of natural increase of the poor may thereby differentially increase, widening in the next generation already large personal income and wealth differences.

Rolling back such a fundamental constraint on human life leads to rearrangements. Family decision making is affected most directly, for the family deals with the periods of economic dependency in the life cycle when the force of mortality is most heavy. Other functions of the family may be affected to a lesser degree: the transmission to heirs of a cultural heritage and the skills for a livelihood. Documentation that mortality change has in fact modified behavior has only recently gotten underway (see for example, Rati Ram and Theodore W. Schultz). The purpose of this paper is to present some evidence on one form of household response to mortality.

The tie between mortality and fertility is a complex one. On the way from being a traditional society to becoming modern, mortality and then fertility are generally observed to decline. This process, called the "demographic transition," remains imprecisely characterized in terms of underlying mechanisms, time dimension and relative magnitudes of change. Conflicting empirical evidence adds to the conceptual ambiguity: cross-sectional data suggests fertility and child mortality are positively related in low income countries (see the author, 1976b), whereas aggregate time-series in these countries show crude death rates falling for three decades before crude birth rates widely decline (see Kuznets).

Changes in general mortality levels are associated across populations with monotonic variation in mortality rates by age. Discussion focuses here on the effect of survival of a mother's own children on her fertility, because the causal relation is direct and obvious, child deaths are a substantial proportion of all deaths, and offspring survival is measured, though unfortunately subject to errors of recall. Concurrent change in child mortality outside of the nuclear family and adult mortality undoubtedly reinforce the incentives to modify reproductive goals, but adult mortality cannot be readily observed in the family and it is difficult to obtain other good proxies for relevant mortality conditions.

*Yale University. This work was partly supported by AID contract otr-1432 and was facilitated by a grant from the Rockefeller Foundation for Economic Demography. I have benefited from the comments of E. A. Hanushek, S. Rosen, and R. Willis.

The paper is organized as follows: Section I reviews how economists have argued that child mortality may affect reproductive behavior, and discusses the unresolved problems of separating biological, behavioral, and simultaneous relations. Though the economic framework yields plausible results, refutable predictions are scarce and empirical work is in order to focus theoretical developments. Section II summarizes aggregate intercountry evidence of fertility responses within a birth cohort to variation in child mortality. Section III reports similar estimates within families based on household survey data for urban Latin America and rural India.

I. Child Mortality and Reproductive Behavior and Motivations

There are four interrelated reasons why one expects fertility and child mortality to be associated: fertility may respond to expected or experienced child (and adult) mortality; child mortality may be influenced by fertility and the proportion of childbearing occurring to women at high risk, for example, the very young, the old, and the poor; both mortality and fertility may be affected by common observable factors, such as education; and finally, both may also be influenced by unobserved factors that generate a correlation between disturbances in equations determining fertility and child mortality.

To distinguish the above sources of covariation, a system of two equations determining fertility and child mortality must be estimated, in all likelihood using simultaneous equation techniques. Without a priori theoretical insights for identifying the two underlying structural relations, or panel data of time-series for couples, it seems premature to extract system estimates from a single cross section identified artificially. For example, only a few percent of the observed variation in child mortality rates across rural Indian families can be traced to a host of standard economic, demographic, or social variables. Furthermore, for use in policy or projection, one wants an estimate of how fertility responds to mortality variation that stems from both economic conditions facing the family, such as those related to family expenditures on food, and those conditions that are exogenous to family resources and market prices, such as public health programs that eradicate smallpox or control malaria.

In estimating the response of reproductive behavior here, the regime of mortality faced by the family is treated in the most simple fashion—predetermined. Such an estimate undoubtedly neglects some systematic though small feedback effect of fertility on child mortality, holding constant several conditioning factors, and may also be slightly biased by residual simultaneous sources of variation in vital rates. Future work will impose identifying restrictions and test for independence of residuals across structural equations (see De-Min Wu).

Empirical estimates of the response of fertility to child mortality also embody both voluntary modification of behavior and involuntary biological processes that constrain reproductive potential. The survival of breast-fed infants can lengthen their mothers' period of sterility following their birth, and thereby delay subsequent births. Under extreme assumptions the biological reduction in births due to the reduction in infant deaths is less than one-third, whereas more realistic exposure parameters for Latin America would suggest a maximum biological response of one-tenth (see S. H. Preston, p. 13). If empirical estimates of the derivative of births with respect to child deaths exceed .1 or at most .2, the excess is likely to arise from voluntary response patterns.

Economists have only recently begun to describe how fertility goals might adjust to the sequential and expected incidence of child mortality. First, leaving aside uncertainty, knowledge that a particular fraction of offspring will die before reaching a specific mature age has two offsetting effects: it increases the cost per survivor, and increases the number of births required to obtain a survivor. If the desired number of survivors is insensitive to or inelastic with respect to their cost and costs are inversely

proportional to survival probabilities, the number of births sought will vary directly with mortality. Alternatively, if parent demands for survivors are cost elastic, reducing the heavy costs of child mortality encourages parents to have more births. However to speak of a "survivor" as the relevant metric for framing parental goals neglects changes in the composition of benefits and opportunity costs as a child ages, and their dependence on family size and child spacing.

Moreover, the resources parents forego to have a child need not be the same across families or within families, even if the prices of relevant market inputs and opportunities for child labor are identical. If the resource intensity of childrearing is viewed by parents as in part a long-term investment in their offspring, an exogenous reduction in mortality encourages more intensive child investments, such as schooling, migration, and more health investments, probably as a substitute for additional children (see Donald O'Hara).

When uncertainty due to mortality is explicitly considered, issues of hedging, insurance, and risk aversion enter; parents may modify their reproductive target in response to uncertainty according to their preferences with respect to family size and the distribution of opportunity costs associated with an excess or shortfall in survivors (see the author, 1969). Research measuring preferences and opportunity costs with respect to family size has progressed slowly and as yet has not dealt directly with how these measures interact with mortality in the determination of reproductive behavior (see for example, J. M. Roberts, R. F. Strand, and E. Burmeister; L. Coombs; P. H. Lindert). The most satisfactory treatment of uncertainty is that developed by Yoram Ben-Porath and Finis Welch, though its empirical application has lagged.

Given the concentration of child mortality in the initial years of life and the relatively long period of childbearing, the need for parents to hedge against child mortality appears low. Sequential decision making permits parents, for the most part, to replace deceased infants rather than bear (or withhold) additional children as a hedge against expected but uncertain future mortality. Adjustment to declines in child mortality can therefore be largely accomplished *ex post*, if parents are able to refrain from replacing offspring who (unexpectedly) survive. This capacity to adjust fertility initially through *replacement* rather than according to *expectations* creates the potential for a short lag between mortality and fertility. In the longer run, as mortality expectations adjust, an entirely new age pattern of reproductive behavior and child investment may emerge, conforming to perceived benefits and costs of birth spacing, family portfolios, and life cycle investment schedules.

A decade may nonetheless elapse between the initial decline in child mortality and the beginning of a decline in replacement fertility in the family formation process. Even then, the change in reproductive behavior may go unnoticed because of the small numerical importance of births to women in their late 30's and 40's. At the aggregate level, changes in age composition, induced by the age pattern of mortality declines, first depress crude birth rates slightly below their age standardized path for about a decade, and then increase them notably thereafter.[1] Crude birth rates 25 years after the onset of the mortality decline may thus rise, even as age standardized birth rates subside. Finally, resource and price constraints facing households may also change, raising or lowering reproductive goals independently of mortality. At this time, therefore, crude birth and death rates cannot directly clarify how fertility is adapting at the family level to child mortality mainly because changes in age composition are often unobserved.

[1]The reversal in crude and age standardized fertility trends are seen in several countries of Latin America and Asia in the mid-1960's, where age-specific vital rates are relatively reliable. Crude death rates, of course, fall much faster than age standardized mortality measures, increasing population growth rates beyond sustainable levels in the first several decades of mortality decline. Hence, the paradox that crude death rates in countries like Taiwan and Puerto Rico are about half of *U.S.* levels today.

In sum, economic logic has not yet described in a refutable form how fertility should be related to child mortality in equilibrium. In addition, the inaccessibility and cost of sufficiently reliable and acceptable birth control methods could, in many parts of the world, introduce another indeterminant innovational lag. The balance of the paper assembles some evidence of the actual empirical relation.

II. Aggregate Evidence on Fertility-Child Mortality Relationships

Widely available aggregate data permit testing of only a few questions about the relation between fertility and mortality. The minimum of data needed to estimate replacement response is the number of children born to comparably situated women who experienced different rates of child mortality. There are large parts of the world where no census information of this nature exists, including the United States.

As an example of the type of research that can be conducted with census data, it is possible to test if the number of living offspring of women of a specific age is roughly constant in a particular year, irregardless of child mortality. The requisite information is available for some eighty countries for rural/urban or total populations ($N = 95$). By observing that the average number of children alive per woman A, equals the number born alive C, times a survival rate $p = 1 - (D/C)$, (where $D = C - A$), the logarithm of C can be regressed on the logarithm of $1/p$ and the calendar year, t, to which the data pertain as in:

$$(1) \quad lnC_i = \alpha_0 + \alpha_1 t_i + \alpha_2 ln\,(1/(1 - (D_i/C_i))) \qquad (i = 1, \ldots, 95)$$

If estimates of α_2 equal one, within age groups of women, the cross-sectional variation in cumulative fertility is simply replacement.

For the estimated equations, the R^2's range from .2 to .3. The point estimates of α_2 (followed by their standard errors) among older women, for whom replacement is more nearly complete, cluster in the vicinity of one: age 30–34, 1.10 (.26); 35–39, 1.06 (.26); 40–49, .98 (.25); 50 or more, .77 (.16) (see the author, 1976b, Table 8.3). The hypothesis that α_2 equals one cannot be rejected; however, these estimates are undoubtedly biased upwards (see the author, 1976b).

Other cross-sectional estimates of this and other models based on regional variation in cumulative fertility and mortality rates suggest fully half of the variation in fertility is offset by differences in child mortality (see the author, 1976b). Unfortunately, there are only nine developing countries with relevant census data for two points in time.[2] Although in no way representative, this handful of cases shows that fertility declines compensated for about one-half of the concurrent declines in cohort child mortality for women aged 40–49.

III. Individual Analysis of Fertility and Own Child Mortality

Individual data provides a richer test of the interaction between fertility and child mortality. Analysis deals here with several representative samples of women age 30–49, with one or more births whose husbands are present, drawn from three Latin American urban surveys from 1964 conducted by Centro Latino-Americano de Demografiá (*CELADE*), and a rural Indian survey from 1970 conducted by the National Council of Applied Economic Research (*NCAER*).[3] Various theories have stressed different factors influencing fertility: these are reviewed elsewhere (see the author, 1976a). With these data, it is

[2]Data from U.N. *Demographic Yearbooks* available repeatedly for Bermuda, Brazil, Cyprus, Fiji, Gilbert and Ellice Islands, Hong Kong, Indonesia, Philippines, and the Solomon Islands.

[3]Survey data was kindly provided by two research centers. The *CELADE* in Santiago, Chile, coordinated under Carmen Miro's direction a series of urban comparative fertility surveys each of about 2,000 women of childbearing age. The Indian Additional Rural Income Survey of about 6,000 rural households was directed by M.T.R. Sarma during 1968–71 at the *NCAER*.

possible to hold constant at least a few price, income, and origin variables in order to assess the partial association between cumulative fertility and cumulative child mortality. Specifically, the education of the husband is included as an income effect. The education of the wife captures not only an income but also a dominant substitution effect, which may be attributed to the opportunity cost of her time in child rearing. Permanent income cannot be measured identically across samples; in urban areas it is the logarithm of household monthly expenditures exclusive of housing, and in rural areas it is total family income. The income variable is not consistently associated with fertility, controlling for husband's education, but it is inversely related to child mortality in all samples. Age of wife is included as a control for life cycle or birth cohort differences. Migrant origins and duration of city residence are held constant in the Latin American samples, while the presence of village health and educational institutions are included as controls in the rural Indian sample to capture local access to health and schooling services. The number of deceased children is normalized as a fraction of those predicted based on the woman's number of births, the age pattern of fertility, and an appropriate life table.[4]

Table 1 shows the results from the regression of children ever born on the normalized child mortality rate and the aforementioned variables. The regression coefficient on mortality and its t-ratio are shown in column (1). Columns (6) and (7) provide means of the fertility and normalized child mortality variables; the derivative of births with respect to deaths is shown in column (3) (see fn. b, Table 1). In all thirteen samples the level of fertility is positively associated with child mortality, and in all cases except women 40–49 in Mexico City, the associations are statistically significant (10 percent level). The derivative of births with respect to child deaths ranges widely, however, from .8–1.4 in Rio de Janeiro, to .4–.8 in San José, to .2–.3 in Mexico City, and in rural India between .3 and .5.

Why should fertility responses to child mortality show such substantial variation across populations? I propose the hypothesis that couples react to their child mortality experience by changing their reproductive performance, to the extent that they are aware of a general downtrend in mortality in their segment of society. The absolute levels of child mortality among women 40–49 are initially similar (not reported) in the three city samples. They declined by 10 percent in Mexico City in the thirteen years spanned by these data (approximately 1945–58), 30 percent in Rio, and 40 percent in San José (col. (7)). Fertility decreased little in Mexico across these birth cohorts, whereas it declined notably in Rio and San José, and offsetting individualistic responses of fertility to own child mortality were substantial in the latter cities but not in the former.

The Indian sample is divided into cultivators working their own land and landless rural laborers, but fertility, mortality and response coefficients are not notably different (5 percent level) between these subsamples. Mortality is greater in India than in urban Latin America, and the secular downtrend across cohorts is moderate after a drop in the early 1950's. The compensatory response derivative of fertility with respect to child mortality increases to about one-half among Indian women at the end of their childbearing years, age 40–49.

To explore how fertility responds to child mortality across economic classes within a society, two income groups are defined (based not on observed income which might be endogenous, but on an instrumental variable prediction of income or expenditures derived from husband age, education, and origins). Column (2) reports the regression coefficient on a dummy variable interacted with the mortality vari-

[4] A. J. Coale and T. J. Trussel birth schedules by age are first scaled down to yield the cumulative fertility reported by each woman in the survey. This imputed flow of births is then subjected to national age-specific mortality rates to obtain an expected number of child deaths. Costa Rican and Mexican life tables are for 1966, the Indian 1961, and lacking a Brazilian table, the one for Colombia in 1965 was substituted as reasonable for health conditions in Rio de Janeiro (see Nathan Keyfitz and Wilhelm Flieger).

TABLE 1—ASSOCIATION BETWEEN CUMULATIVE FERTILITY AND CHILD MORTALITY IN SELECTED SAMPLES OF HOUSEHOLDS: URBAN LATIN AMERICA 1964 AND RURAL INDIA 1970[a]

	Coefficient on Child Mortality in Fertility Equation[d]		Derivative of Births with respect to Child Deaths at Sample Means[b]			Overall Variable Means[e]		Ratio of Upper to Lower Income Class Variable Means	
	Overall (1)	Class Difference (2)	Overall (3)	Upper Class (4)	Lower Class (5)	Fertility (6)	Mortality (7)	Fertility (8)	Mortality (9)
Rio de Janeiro, Brazil, 1964									
30–34	.304	−.113	.752	.899	.650	3.14	.488	.746	.250
	(3.94)	(.43)	(315)	(122)	(193)	(1.92)	(1.31)		
35–39	.617	.295	1.35	1.68	1.17	3.21	.370	.697	.686
	(4.78)	(1.16)	(279)	(120)	(159)	(2.20)	(.973)		
40–49	.541	.832	.941	1.10[c]	.825[c]	3.76	.681	.768	.604
	(4.95)	(3.87)	(325)	(124)	(201)	(2.70)	(1.32)		
San José, Costa Rica, 1964									
30–34	.206	.0554	.561	2.95	.508	4.03	.539	.761	.221
	(2.33)	(.19)	(268)	(99)	(169)	(2.06)	(1.37)		
35–39	.429	−.341	.827	1.18	.771	5.04	.867	.668	.446
	(3.67)	(1.15)	(240)	(80)	(160)	(2.90)	(1.54)		
40–49	.212	−.462	.397	.510	.362	5.35	.936	.709	.398
	(1.85)	(1.42)	(287)	(81)	(206)	(3.24)	(1.64)		
Mexico City, D.F., Mexico, 1964									
30–34	.123	.357	.252	.322[c]	.221[c]	4.48	.939	.718	.635
	(1.55)	(2.30)	(2.95)	(116)	(179)	(2.27)	(1.60)		
35–39	.112	.0666	.184	.217	.166	5.43	.925	.799	.438
	(1.97)	(.78)	(239)	(98)	(141)	(2.86)	(1.57)		
40–49	.121	−.218	.174	.224	.152	5.57	1.07	.696	.447
	(1.19)	(1.02)	(348)	(136)	(212)	(3.27)	(1.70)		
India Rural *ARIS*, 1970									
30–39									
Land Owners	.265	−.0897	.246	.240	.252	4.22	.351	1.06	1.09
	(3.08)	(.55)	(832)	(376)	(456)	(1.80)	(.708)		
Landless	.408	−.920	.382	.344[c]	.407[c]	4.17	.295	1.17	.703
	(2.97)	(2.79)	(349)	(133)	(216)	(1.82)	(.673)		
40–49									
Land Owners	.768	−.415	.523	.502	.549	5.04	.320	1.10	.634
	(5.37)	(1.40)	(621)	(324)	(297)	(2.15)	(.582)		
Landless	.782	−.0163	.539	.518	.559	4.88	.426	1.29	.774
	(4.36)	(.97)	(277)	(134)	(143)	(2.22)	(.734)		

[a]Derivation of columns described in text. Underlying regression results obtainable from author.

[b]Derivative $dC/dD = \beta(C/D)/(\bar{C} + \beta(C/D)\,(\overline{D/C}))$ where β is the regression coefficient of the normalized child death rate, (C/D) is the reciprocal of the sample mean expected child death rate, $\bar{C}$ and $(\overline{D/C})$ the sample means of fertility and actual child mortality. Sample size is shown in parentheses.

[c]If distinct regression coefficients are estimated for the two income classes in those cases where the t-ratio exceeds 1.6, the derivative of births with respect to child death is .0663 and 1.25 for upper and lower class samples age 40–49 in Rio, −.231 and .474 for age 30–34 in Mexico City, and 1.13 and .209 for age 30–39 in the rural Indian landless sample, by income class.

[d]The t-ratios are shown in parentheses.

[e]The standard deviations are shown in parentheses.

able for the lower income group and its t-ratio. The null hypothesis of coefficient equality is rejected (10 percent level) in only three out of thirteen cases, suggesting a common estimated coefficient may exist for both income classes. It should be stressed that given the small size of these samples and the relative infrequency of child mortality, little confidence can be placed on differences between these groups, except as a source of working hypotheses.

When derivatives are evaluated at subsample means using the common regression coefficient across income groups, the response patterns are similar for India but quite different in Latin America (cols. (4) and (5)), largely because the urban fertility and child mortality are substantially lower at higher income levels. In addition, the large declines in child mortality in Rio and San José that were noted earlier appear to have most benefited the upper income classes (col. (9)). Consistent with my hypothesis, the derivative response of fertility with respect to mortality is also larger for these upper income classes. Recall also that household income and child mortality is negatively correlated in all samples. This may indicate that the widely accepted view that economic development plays only a minor role in the remarkable mortality transition in low income countries (see Stolnitz) needs reevaluation. Analysis is needed on how improvements in the economic environment of the family and its behavior influence prospects for child survival, and in turn impinge on fertility and other forms of human and physical capital investment within the family. In other words, the next step is to estimate the full structural equation model determining both fertility and child mortality at the family level.

IV. Conclusions

Across samples of urban and rural households in Latin America and India, statistically significant associations are reported between cumulative fertility and cumulative child mortality, holding constant age, education, income, and origins. Individual reproductive responses to child mortality increase to fully compensating levels only in those populations where child survival has markedly improved. Aggregate trends as well as individual child survival experience should be examined jointly in future efforts to understand individual reproductive behavior in low income countries. The economic determinants and consequences of mortality now warrant more study, given the magnitude of recorded change in life expectancy and our nearly complete ignorance of who has benefited by this significant process during economic development and why.

REFERENCES

Y. Ben-Porath and F. Welch, "Chance, Child Traits, and the Choice of Family Size," The Rand Corp., R-1117-NIH/RF, 1972.

A. J. Coale and T. J. Trussell, "Model Fertility Schedules," *Population Index*, Apr. 1974, *40*, 185 257.

L. Coombs, *Are Cross Cultural Preference Comparisons Possible? A Measurement Theoretic Approach*, IUSSP paper no. 5, Liege, Belgium 1976.

Nathan Keyfitz and Wilhelm Flieger, *World Population: An Analysis of Vital Data*, Chicago 1968.

S. Kuznets, "Recent Population Trends in Less Developed Countries and Implications for Internal Income Inequality," disc. paper no. 261, Yale Univ., May 1977.

P. H. Lindert, "Child Costs and Economic Development," in *Population and Economic Change in Less Developed Countries*, Universities-Nat. Bur. Econ. Res. conference series, Philadelphia, Sept. 1976.

D. J. O'Hara, "Changes in Mortality Levels and Family Decisions Regarding Children," The Rand Corp., R-914-RF, Feb. 1972.

S. H. Preston, "Introduction," *Seminar on*

Infant Mortality in Relaion to the Level of Fertility, CICRED, Paris 1975.

R. Ram and T. W. Schultz, "Some Economic Implications of Increases in Life Span with Special Reference to India," mimeo., Univ. Chicago, June 1977.

J. M. Roberts, R. F. Strand, and E. Burmeister, "Preferential Pattern Analysis," in P. Kay, ed., *Explorations in Mathematical Anthropology*, Cambridge, Mass. 1971.

T. P. Schultz, "An Economic Model of Family Planning and Fertility," *J. Polit. Econ.* Mar./Apr. 1969, *77*, 153–80.

———, (1976a) "Determinants of Fertility: A Micro-economic Model of Choice," in Ansley J. Coale, ed., *Economic Factors in Population Growth*, New York 1976.

———, (1976b) "Interrelationships Between Mortality and Fertility," in Ronald G. Ridker, ed., *Population and Development*, Baltimore 1976.

G. J. Stolnitz, "International Mortality Trends," in *The Population Debate*, New York 1975.

D.-M. Wu, "Alternative Tests of Independence Between Stochastic Regressions and their Disturbances," *Econometrica*, July 1973, *41*, 733–50.

United Nations, *Demographic Yearbook*, New York, various years.

Altruism as an Outcome of Social Interaction

By MORDECAI KURZ*

In an earlier paper (1977) I presented a model of altruistic behavior which has only a trivial equilibrium in a single period economy but has a nontrivial set of equilibria when the economy is repeated. The essence of such equilibria is the establishment of reciprocal altruism over time due to the repetitive nature of the economy. The main problem which was left open in that paper is the great multitude of altruistic equilibria; this problem, is, in fact, common to all models of equilibria in the supergame (see for example, Robert Aumann and Lloyd Shapley; A. Rubinstein). The present paper explores various social concepts and mechanisms which would narrow down the set of equilibria.

I. Review of the Model of Altruistic Behavior

Following the earlier paper, consider an economy $E = ((\alpha_1, \omega_1), (\alpha_2, \omega_2), \ldots, (\alpha_n, \omega_n))$ composed of n individuals each with an endowment $\omega_i \in R^l_+$ and preference ordering α_i defined over the consumption set $X_i \subset R^l$ and is represented by the continuous utility function $u_i(x_i)$, $x_i \in X_i$. I assume that all commodities can be defined as desirable, so that $u_i(x_i + d_i) \geq u_i(x_i)$ for all $d_i \in R^l_+$. The consumption set X_i may contain both the usual economic commodities and commodities which are not normally traded such as the degree of affection between two persons, the amount of help an individual will give to others in distress, and the degree of orderly behavior in public places. The only requirement is that a commodity be divisible, identifiable, and all individuals know the amounts available, given and received. Let

z_i = vector of goods which i keeps for himself

g_i = vector of goods which i gives to others

y_{ij} = vector of goods which individual i gives to j ($y_{ij} \geq 0$, $y_{ii} = 0$ all i)

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{in}) \text{ and } g_i = \sum_{j=1}^{n} y_{ij}$$

$r_i = \sum_{k=1}^{n} y_{ki}$ = vector of goods which individual i receives from others

$N = \{1, 2, \ldots, n\}$ = the set of individuals participating.

The game of altruism is a noncooperative game which can be described very simply. Each individual player can select an action $y_i \in R^{ln}_+$, $i = 1,2,\ldots,n$ which results in his giving a vector y_i to others, retaining for himself the vector $(\omega_i - g_i)$, receiving from others a vector r_i and therefore consuming $x_i = \omega_i - g_i + r_i$. Thus in a single play, a strategy n-tuple $y = (y_1, y_2, \ldots, y_n)$ results in a payoff vector $(u_1(x_1), u_2(x_2), \ldots, u_n(x_n))$. In my 1977 paper I show that if this game is played a finite number of times $t = 1,2,\ldots,T$ then the only Nash equilibrium is $y^t_i = 0$ for all t and $x^t_i = \omega^t_i$. However, if the game is repeated infinite number of times, the endowment vectors ω^t_i remain stationary (i.e., $\omega^t_i = \omega_i$ all t) and if the supergame utility U is taken to be the mean utility over all outcomes, then the set of

*Stanford University. This work was carried out when I was a Guggenheim Fellow. The work was also partially supported by National Science Foundation grant SOC75-21820 at the Institute for Mathematical Studies in the Social Sciences (*IMSSS*), Stanford University. I benefited from extensive discussions on the subject with Robert Aumann and Reinhard Selten. I also wish to thank Kenneth Arrow, Peter Hammond, Eric Maskin, and members of the *IMSSS* Seminar on income distribution for valuable comments.

Nash equilibria generates simply the set of all individually rational allocations. Thus if σ^* is a Nash equilibrium in the supergame, then $U_i(\sigma^*) \geqq u_i(\omega_i)$ for all i. Moreover consider any stationary allocation $x^t = x$, $t = 1,2, \ldots$ with $x = (x_1,x_2, \ldots ,x_n)$; if $u_i(x_i) \geqq u_i(\omega_i)$ for all i, then this stationary allocation can be sustained by a Nash equilibrium in the supergame.

In the two-person case one notes in Figure 1a that any allocation in the area to the northeast of the point $u_1(\omega_1)$, $u_i(\omega_2)$ on or below the frontier can be obtained as a Nash equilibrium in the supergame.

It may be of benefit to consider how the present model of altruism differs from notions of altruism discussed in the biological literature (see, for example, W. D. Hamilton; Robert Trivers; Ilan Eshel; and others). In my view the fundamental distinction is to be found in the underlying concept of rationality used in the model involved. In the natural universe, the survival of some species may depend upon their adopting behavior of altruistic reciprocity but there is nothing in these natural phenomena to suggest that an individual member of a species "calculates" that its optimal behavior is to be altruistic. In fact, most of the examples given suggest that animals develop a very strong resistance mechanism to deviations from their mode of behavior since they have been conditioned to act the way they do.

My model of altruistic behavior allows

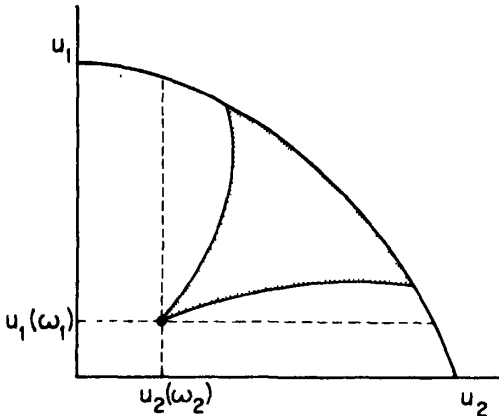



FIGURE 1a

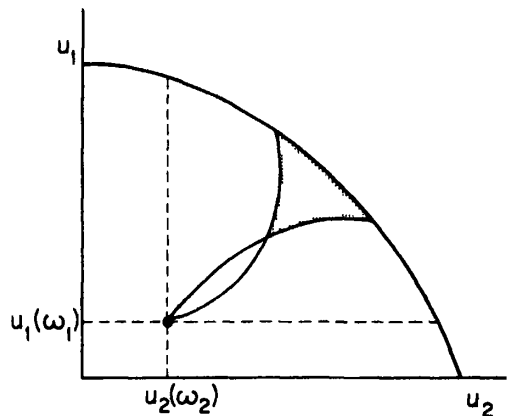



FIGURE 1b

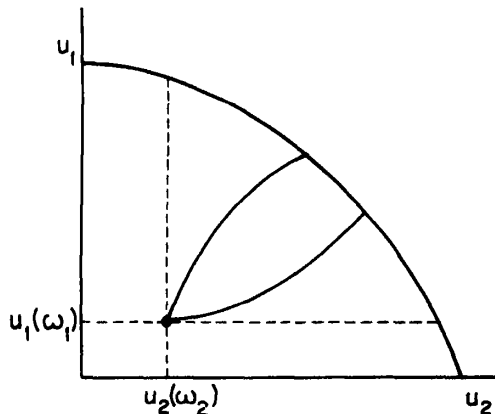



FIGURE 1c

each player the individual choice of acting altruistically or selfishly. In an altruistic equilibrium all individuals act altruistically since it is in their individualistic best interest to act that way. Thus in my altruistic equilibrium it is also *individually* rational to act altruistically and in that sense individual rationality reinforces the social order.

To argue that human beings do not act rationally and, like other members of the natural universe, only respond in a conditioned manner, raises a few fundamental issues in social theory. To clarify my position let me state what kind of rationality I presume only humans to possess and thus provide a basic delineation between my theory and that of the biological literature: 1) Each individual is assumed to have an intertemporal utility function, thus an allocation decision must consider both the immediate and *future needs* and 2) Every individual is capable of predicting how his actions will influence the reactions of other participants and he will take this prediction into account when making his final choice. It appears to me that to ascribe this kind of rationality to individual members of all species would be unwarranted.

Before proceeding with the investigation I note that I often assume that the utility functions and the endowments of all the players remain stationary; by concentrating on stationary strategies and equilibria I shall in fact be covering most of the interesting cases and illustrating the main points.

II. Discounting

In my earlier paper (1977) I consider the supergame utility function U to be the mean utility over all plays. Consider now the effect of introducing discounting. This means that the supergame utility of i is defined by[1]

$$(1) \qquad U_i(\sigma) = \sum_{k=0}^{\infty} \left(\frac{1}{1+\rho_i} \right)^k u_i(x_i^k)$$

[1] I do not write limits in (1) since by monotonicity $u_i(x_i^k) \le u_i(\omega) \le \beta < \infty$ uniformly in k and i and where $\omega = \sum_{i=1}^{n} \omega_i$. This means that as long as $\rho_i > 0$ the sum in (1) converges and $U_i(\sigma)$ is well defined.

Suppose that σ^* is a stationary strategy which is a Nash equilibrium in the supergame and let $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ be the associated stationary supergame allocation. If at time $k = 0$ player j defects and sets $g_j^0 = 0$ and assuming complete retaliation by all the other players he should expect to end up with the following consumption flow $(\omega_j + r_j^*, \omega_j, \omega_j, \ldots)$. For σ^* to be a Nash equilibrium, that player will not defect since

$$(2) \qquad \frac{u_j(x_j^*) - u_j(\omega_j)}{\rho_j} \ge u_j(\omega_j + r_j^*) - u_j(x_j^*)$$

The interpretation of (2) is rather direct: the expression $(u_j(x_j^*) - u_j(\omega_j))/\rho_j$ is the capitalized net present value from time 0 to infinity of compliance with σ^* while $[u_j(\omega_j + r_j^*) - u_j(x_j^*)]$ is the net gain from defecting.

Equation (2) defines a set of stationary allocations which would constitute Nash equilibria in the supergame. For the case of $n = 2$,

$$x_1^* = \omega_1 - g_1^* + r_1^*$$
$$x_2^* = \omega_2 - g_2^* + r_2^*$$

thus $g_1^* = r_2^*$ and $r_1^* = g_2^*$. It follows that we can write (2) as

$$(3a) \quad u_1(\omega_1 - r_2^* + r_1^*) - u_1(\omega_1) \ge \rho_1[u_1(\omega_1 + r_1^*) - u_1(\omega_1 - r_2^* + r_1^*)]$$

$$(3b) \quad u_2(\omega_2 - r_1^* + r_2^*) - u_2(\omega_2) \ge \rho_2[u_2(\omega_2 + r_2^*) - u_2(\omega_2 - r_1^* + r_2^*)]$$

Hence we can think of the pair (r_1^*, r_2^*) as defining the utility levels which are feasible for any given pair of discount rates (ρ_1, ρ_2). In Figure 1a the shaded area indicates all the possible Nash equilibrium allocations achieved for small (ρ_1, ρ_2) while in Figure 1b the set (which also includes the point $(u_1(\omega_1), u_2(\omega_2))$) is delineated for large (ρ_1, ρ_2). In the limit, as either ρ_1 or ρ_2 become very large, the situation of Figure 1c occurs in which the only equilibrium is the trivial one at which $x_i^* = \omega_i$ all i.

The intuitive meaning of these results is clear: as the discount rates rise, the myopic incentive to defect and take advantage of

the temporary gain from defection increases thus narrowing the set of allocations which would constitute an equilibrium. As either ρ_1 or ρ_2 tends to become very large the only possible equilibrium is the trivial one in which all altruistic behavior vanishes.

III. Perfect Equilibrium

For games in extensive form it has been widely argued that "Nash equilibrium" is an inadequate concept since it does not possess any optimality properties at points which were not called for by the equilibrium strategies (see R. Selten; Aumann and Shapley). The notion of "perfectness" of equilibrium proposed by Selten was intended to remedy the situation and therefore it could be valuable to consider the set of perfect equilibria in the game of altruism proposed above.

To clarify the notion of perfectness as applied to the game at hand denote by G the single play game which I described earlier and by G^* the supergame. If the game is played for $T - 1$ periods then I denote the continuation game from T on by G_T^*. Note that any strategy σ for G^* induces a strategy σ_T for G_T^*. *A strategy σ^* is said to be a perfect equilibrium if its induced strategies σ_T^* for all $T = 0, 1, 2, \ldots$, are Nash equilibria in all the continuation supergames.*

The intuitive sense in which the set of perfect equilibria can be perceived to be smaller than the set of Nash equilibria is that if a player deviates from a set of actions prescribed by a Nash equilibrium then the retaliation against him may be damaging to the retaliating players. Hence, if a player deviates at $T - 1$, the induced strategy may be unacceptable as a Nash equilibrium in the supergame G_T^* since it may inflict a great deal of harm on the punishing players. This means that when play resumes at time T the players may prefer, each one individually, to discontinue any such punishing strategy which does too much harm to themselves. In spite of this intuitive sense, it does not apply to the game of altruism and for this game I present the basic result:

THEOREM 1: *Regardless of the form of the supergame utility functions U_i, the set of payoffs associated with the set of all Nash equilibria in G^* coincides with the set of payoffs associated with the set of perfect equilibria in G^*.*

I remark that for the case in which the utility function is the mean over the utilities of the individual plays, both Aumann and Shapley and Rubinstein have shown that the theorem above is true for all games. What is stated here is that for the game of altruism the theorem is true for all supergame utility functions. The proof of Theorem 1 is omitted.

The reader may note that the set of perfect equilibria is smaller than the set of Nash equilibria whenever there is a way in which each individual player has available to him an action in which he can punish the deviator at some cost to himself. The set of equilibria will then depend upon the force of the punishment relative to the cost of punishment. If the cost of punishment is too high, each individual player may not wish to persist in punishment since its continuation injures him too much (see Aumann and Shapley). In the game of altruism the punishing capability of each individual player is limited. Moreover, when all are supposed to play selfishly, a player has no punishing capability at all since attempting to deviate from selfish play, when all the other players continue to play selfishly, will only worsen his situation. I thus conclude that due to the absence of the player's punishment capability, the notion of "perfectness" does not help in narrowing down the set of altruistic equilibria.

IV. Limited Rationality and Incomplete Information

In the broad social context of the analysis here one can associate with every Nash equilibrium an ethical system which is reflected by specific norms of Conduct and Belief. With this in mind authors like P. Hammond and Edmund Phelps argue that one needs an additional structure in

order to narrow down the set of social equilibria. For this reason I will explore in the remainder of this paper an additional structure based on limited rationality and incomplete information. The discussion here is strictly preliminary and exploratory in nature with the full development postponed to a later paper.

I propose to think of the n participants as engaging in a pregame search procedure from which they may learn the unknown true state of the world which is defined by a set of ethical standards associated with the equilibrium which will prevail. This will tell them which strategy to play in the game of altruism. Thus let Σ be the set of all Nash equilibria in the game of altruism; each participant forms his probability beliefs on Σ. These beliefs will induce an optimal action where the definitions of "action" and "optimality" are as presented below. Given the observed actions each individual learns something about the state of the world and thus revises and updates his probability beliefs. An "equilibrium" of this pregame procedure will consist of a set of beliefs and a sequence of actions which converge to a point in Σ. As the reader may note below the deeper problem is the question of the convergence of the procedure and this is *not* examined here. What I am exploring here is some characterization of that family of procedures for which an equilibrium of the pregame induces the selection of a *Pareto optimal* Nash equilibrium in the game of altruism. I now turn to the formal statement of the above.

I start with the probability density of individual k at step t of the procedure which is denoted by $R_k^t(\sigma \mid a^1, a^2, \ldots, a^{t-1})$ where $\sigma \in \Sigma$ and $a^\tau = (a_1^\tau, a_2^\tau, \ldots, a_n^\tau)$ is a vector of actions a_k^τ taken by individual k at step τ. The function $R_k^t(\sigma \mid a^1, a^2, \ldots, a^{t-1})$ incorporates the learning process of individual k resulting in a continuous update of his probability beliefs. The key question is what does an action mean and in what sense is it optimal.

Individual k knows that the only point at which the procedure stops is when the actions of all individuals combined constitute a point in Σ. Thus in selecting his action a_k^t our individual assumes that either there will be an equilibrium established or we go back to square 1 and start all over again. Normalizing $u_k(\omega_k) = 0$ for all k, at every step t each individual forms his expected utility V_k^t as

$$(4) \quad V_k^t(\sigma, a^1, a^2, \ldots, a^{t-1}) = u_k(\sigma)\, R_k^t(\sigma \mid a^1, a^2, \ldots, a^{t-1})$$

Now define $\bar{\sigma}^k(t)$ as that equilibrium which would maximize $V_k^t(\sigma \mid a^1, a^2, \ldots, a^{t-1})$ and individual k selects his action a_k^t to be the kth component of the strategy $\bar{\sigma}^k(t)$, i.e., $a_k^t = \bar{\sigma}_k^k(t)$. I thus define a_k^t to be an optimal action by individual k at step t in this sense.

Obviously the individual optimizer selects his action so as to maximize his expected utility and in this sense he is selfish. Yet he can be regarded as acting altruistically in the sense that he does not attempt to manipulate the society or "bargain" with the rest of the players. The individual should be viewed as seeking to learn the true social ethic in order to behave accordingly rather than attempt to manipulate it.

An *equilibrium* in the pregame procedure is a set of updating functions $R_k^t(\sigma \mid \cdot)$ all k and all t and a sequence of actions $(a^1, a^2, a^3, \ldots,)$ such that $a^t \to a$ and $a = \sigma$ is a strategy $\sigma \in \Sigma$, and a_k^t is an optimal action of individual k at step t for all k and all t.

Example 1:

Consider the following single period matrix game:

		L	R
Player 1	U	(1,1)	(3,0)
	D	(0,3)	(5,5)

The only pure strategy Nash equilibria in the supergame are (U,L) and (D,R). Now define $R_k^t(\sigma \mid \cdot)$ on pure strategies only in the following way: since $R_1^t = (P_1^t(U,L), P_1^t(D,R))$ with $P_1^t(U,L) + P_1^t(D,R) = 1$, let

$$P_1^t(U,L) = \begin{cases} 1/2 & \text{initially} \\ 1 & \text{if the previous action of player 2 was } L \\ 0 & \text{if the previous action of player 2 was } R \end{cases}$$

Define R_2^t symmetrically. Since for both players the payoff at (D,R) is larger than at (U,L) their optimal choices under the initial probabilities are to select (D,R). But the above updating functions with the sequence of resulting actions (D,R), (D,R), (D,R), ... , constitute an equilibrium as defined above.

Note however that if the initial probabilities had been 1/10 for both players, then the equilibrium would have been associated with the inferior outcome (U,L), (U,L), (U,L), ...

Example 2:

Consider the following single period matrix game:

		M_1	M_2	M_3
	L_1	(1,1)	(1,1)	(1,1)
Player 1	L_2	(1,1)	(5,2)	(0,6)
	L_3	(1,1)	(6,0)	(2,5)

The three pure strategy Nash equilibria in the supergame are[2] (L_1, M_1), (L_2, M_2), (L_3, M_3).

Now consider the following updating functions R_1^t which are defined on the pure strategies only: equiprobable (1/3) initial outcomes and

$$P_1^t(L_1,M_1) = \begin{cases} 0 & \text{if the previous action of player 2 was } M_2 \text{ or } M_3 \\ 1 & \text{if the previous action of player 2 was } M_1 \end{cases}$$

$$P_1^t(L_2,M_2) = \begin{cases} 9/10 & \text{if the previous action of player 2 was } M_2 \\ 1/10 & \text{if the previous action of player 2 was } M_3 \end{cases}$$

[2]Obviously (L_1,M_2), (L_1,M_3), (L_2,M_1) and (L_3, M_1) would also constitute equilibria with the (1,1) payoff.

$$P_1^t(L_3,M_3) = \begin{cases} 1/10 & \text{if the previous action of player 2 was } M_2 \\ 9/10 & \text{if the previous action of player 2 was } M_3 \end{cases}$$

If player 2 has a symmetric set of beliefs then the optimal choice in the first round is clearly (L_2,M_3). The optimal choice in the second round is (L_3,M_2) and one can see immediately that the optimal choices with these update functions would result in the following sequence of actions: (L_2,M_3), (L_3,M_2), (L_2,M_3), (L_3,M_2), etc. not converging to a stationary Nash equilibrium.

Consider however the effect of changing the initial probabilities of player 1 from (1/3,1/3,1/3) to (1/10,1/10,8/10). The reader may observe that the optimal sequence becomes (L_3,M_3), (L_3,M_3), (L_3, M_3), ... , etc. which, with the new update functions, does constitute an equilibrium in the pregame procedure.

Example 3:

The matrix is the same as in example 2 and the updating functions of the players are as follows: the initial probabilities for both players and $P_1^t(L_1,M_1)$ remain the same, and

$$P_1^t(L_2,M_2) = \begin{cases} 5/6 & \text{if the previous action of player 2 was } M_3 \\ 1/6 & \text{if the previous action of player 2 was } M_2 \end{cases}$$

$$P_1^t(L_3,M_3) = \begin{cases} 1/6 & \text{if the previous action of player 2 was } M_3 \\ 5/6 & \text{if the previous action of player 2 was } M_2 \end{cases}$$

while for the second player

$$P_2^t(L_1,M_1) = \begin{cases} 0 & \text{if the previous action of player 1 was } L_2 \text{ or } L_3 \\ 1 & \text{if the previous action of player 1 was } L_1 \end{cases}$$

$$P_2^t(L_2,M_2) = \begin{cases} 2/3 & \text{if the previous action of player 1 was } L_2 \\ 1/3 & \text{if the previous action of player 1 was } L_3 \end{cases}$$

$$P_2^t(L_3, M_3) = \begin{cases} 1/3 & \text{if previous action of player 1 was } L_2 \\ 2/3 & \text{if previous action of player 1 was } L_3 \end{cases}$$

In this case the optimal sequence of actions is $(L_2, M_3), (L_2, M_3), (L_2, M_3), \ldots$, which is a convergent sequence but is not a Nash equilibrium. This means that these update functions with the resulting sequence of optimal actions do not constitute an equilibrium in the pregame procedure.

With these examples in mind consider the family of all pregame procedures which have the following property denoted by D (dominance): *For any* σ *and* $\bar{\sigma}$ *in* Σ, *if* $u_k(\sigma) > u_k(\bar{\sigma})$ *for* $k = 1,2, \ldots, n$, *then* $R_k^t(\sigma \mid a^1, a^2, \ldots, a^{t-1}) > R_k^t(\bar{\sigma} \mid a^1, a^2, \ldots, a^{t-1})$ *for* $k = 1,2, \ldots, n$ *and for any set of actions* $(a^1, a^2, \ldots, a^{t-1})$. I then have

THEOREM 2: *For any pregame procedure with property* D, *an equilibrium in the procedure converges to a Nash equilibrium* $\sigma^* \in \Sigma$ *which is Pareto optimal.*

The result thus obtained gives at least an initial suggestive answer to the question with which I started: updating procedures which have an equilibrium must converge to Pareto optimal Nash equilibrium if property D is satisfied. This result should be taken only as a tentative start with a great deal of work to follow.

REFERENCES

R. J. Aumann, "Acceptable Points in General Cooperative n-person Games," in A. W. Tucker and R. D. Luce, ed., *Contributions to the Theory of Games*, Vol. IV, Annals Math. Stud., No. 40, Princeton 1959, 287–324.

——— **and L. Shapley,** "Long Term Competition—A Game Theoretic Analysis," unpublished manuscript, 1976.

G. S. Becker, "Altruism, Egoism and Genetic Fitness: Economics and Sociobiology," *J. Econ. Lit.*, Sept. 1976, *14*, 817–26.

Gerard Debreu, *Theory of Value*, New York 1959.

Ilan Eshel, "On the Neighbor Effect and the Evolution of Altruistic Traits," *J. Theoret. Popul. Biology*, 1972, *3*, 258–77.

W. D. Hamilton, "The Genetical Evolution of Social Behavior," *J. Theoret. Biology*, 1964, *7*, Part I, 1–16, Part II, 17–52.

P. Hammond, "Charity: Altruism or Cooperative Egoism," in Edmund S. Phelps, ed., *Altruism, Morality and Economic Theory*, New York 1975, 115–31.

M. Kurz, "Equilibrium in a Finite Sequence of Markets with Transaction Cost," *Econometrica*, Jan. 1974, *42*, 1–20.

———, "Altruistic Equilibrium," in Bela Balassa and Richard Nelson, eds., *Economic Progress, Private Values, and Public Policy: Essays in Honor of William Fellner*, Amsterdam 1977, 177–200.

E. Phelps, "The Indeterminancy of Game-Equilibrium Growth In the Absence of An Ethic" in his *Altruism, Morality and Economic Theory*, New York 1975, 87–105.

R. Radner, "Existence of Equilibrium of Plans, Prices, and Price Expectations in a Sequence of Markets," *Econometrica*, Mar. 1972, *40*, 289–304.

A. Rubinstein, "Equilibrium in Supergames," res. memo. no. 25, Center Res. Math. Econ. Game Theory, Hebrew Univ., May 1977.

R. Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *Int. J. Game Theory*, No. 1, 1975, *4*, 25–55.

R. L. Trivers, "The Evolution of Reciprocal Altruism," *Quart. Rev. Biology*, Mar. 1971, *46*, 35–57.

Bayesian Decision Theory and Utilitarian Ethics

By JOHN C. HARSANYI*

One of the great intellectual achievements of the twentieth century is the Bayesian theory of rational behavior under risk and uncertainty. Many economists, however, are still unaware of how strong the case really is for Bayesian theory, and many more fail to appreciate the far-reaching implications the Bayesian concept of rationality has for ethics and welfare economics. The purpose of this paper is to argue that the Bayesian rationality postulates are absolutely inescapable criteria of rationality for policy decisions; and to point out that these Bayesian rationality postulates, together with a hardly controversial Pareto optimality requirement, entail *utilitarian ethics* as a matter of mathematical necessity.

I. The Bayesian Rationality Postulates

In discussing the criteria for rational behavior, I will distinguish behavior under certainty, under risk, and under uncertainty. Certainty obtains when we can predict the actual outcome of any action we can take. Risk obtains when we know at least the objective probabilities associated with alternative possible outcomes. Finally, uncertainty obtains when even these objective probabilities are partly or wholly unknown to us (or are possibly even undefined).

In the case of certainty, I will assume that each individual chooses among various alternative situations, where each situation is characterized by finitely many economic and noneconomic variables, such as his holdings of different commodities, including money, his health, his social position, his social relationships, etc., as well as by similar economic and noneconomic variables affecting other individuals in the society. Thus, mathematically, any situation can be regarded as a point in a finite-dimensional (say, r-dimensional) Euclidean space E^r.

In the case of risk and of uncertainty, an individual's choices can be modeled as choices among different lotteries whose "prizes" are situations, that is, points in E^r. A lottery can be described as

$$(1) \quad L = (A_1 \mid e_1, \ldots, A_k \mid e_k, \ldots, A_K \mid e_K)$$

indicating that this lottery L will yield prizes $A_1, \ldots, A_K$, depending on which one of K mutually exclusive and exhaustive events $e_1, \ldots, e_K$ occurs. These events $e_1, \ldots, e_K$ will be called conditioning events. Mathematically, any event e_k ($k = 1, \ldots, K$) can be regarded as a measurable subset of the space Ω of all possible "states of the world." A lottery L will be called a risky or an uncertain lottery, depending on whether the decision maker does or does not know the objective probability $p_k = Prob\ (e_k)$ associated with every event e_k ($k = 1, \ldots, K$) used in this lottery L.

In analyzing the behavior of any individual i ($i = 1, \ldots, n$), strict preference by him will be denoted by $>_i$ and nonstrict preference including indifference, or equivalence, by $\geqq_i$.

Rational behavior by a given individual i under certainty can be characterized by two rationality postulates:

1. *Complete preordering.* The nonstrict preferences of this individual i establish a complete preordering over the space E^r of all possible situations (or over some suitable closed subset of E^r).

2. *Continuity.* Suppose that the sequence $A_1, A_2, \ldots$, of situations converges to a particular situation A_0, and that another sequence $B_1, B_2, \ldots$, of situations converges to B_0, with $A_k \geqq_i B_k$ for all k. Then, $A_0 \geqq_i B_0$.

*Professor of business administration and of economics, University of California-Berkeley. I wish to thank the National Science Foundation for supporting this research through grant SOC77-06394 to the Center for Research in Management Science, University of California-Berkeley.

For convenience I will call these two postulates the *basic utility axioms*. Using these two axioms, we can characterize rational behavior as follows.

THEOREM 1: *Utility Maximization. If an individual's preferences satisfy the two basic utility axioms, then his behavior will be equivalent to maximizing a well-defined (ordinal) utility function.*[1] (For proof, see Gerard Debreu, pp. 55-59.)

To characterize rational behavior under risk and under uncertainty, we need two additional rationality postulates:

3. *Probabilistic equivalence*. Let

$$L = (A_1 \mid e_1, \ldots, A_K \mid e_K) \quad (2)$$

and

$$L^* = (A_1 \mid e_1^*, \ldots, A_K \mid e_K^*)$$

and suppose that the decision maker knows the objective probabilities associated with events $e_1, \ldots, e_K$ as well as with events $e_1^*, \ldots, e_K^*$, and knows that these probabilities satisfy

$$Prob(e_k) = Prob(e_k^*) \quad \text{for } k = 1, \ldots, K \quad (3)$$

Then he will be indifferent between lotteries L and L^*.

In other words, a rational individual will be indifferent between two risky lotteries if these yield him the same prizes with the same probabilities—even if the two lotteries use quite different physical processes to generate these possibilities. Note that this postulate implies von Neumann and Morgenstern's postulate on compound lotteries: that a rational individual will be indifferent between a two-stage lottery and a one-stage lottery if both offer the same prizes with the same probabilities.

[1]For some purposes it may be desirable to define rational behavior without requiring that it should satisfy the continuity postulate (postulate 2). It can be shown that if a given individual's preferences satisfy at least the complete preordering postulate (postulate 1), then his behavior will be equivalent to lexicographically maximizing a certain utility vector.

4. *Sure-thing principle*. Suppose that $A_k^* >_i A_k$ for $k = 1, \ldots, K$. Then

$$(A_1^* \mid e_1, \ldots, A_K^* \mid e_K) \geq_i (A_1 \mid e_1, \ldots, A_K \mid e_K) \quad (4)$$

In other words, other things being equal, a rational individual will not prefer a lottery yielding less desirable prizes over a lottery yielding more desirable prizes. Note that the sure-thing principle is essentially identical with the game-theoretical principle that a rational individual will avoid using any (weakly or strongly) dominated strategy.

Obviously, both postulates 3 and 4 are extremely compelling rationality requirements. But they are subject to two qualifications.

(i) Both postulates presuppose that the utility $U(A_k)$ of any prize A_k is independent of its conditioning event e_k. This requirement can always be satisfied by suitable definition of the prizes. For example, the utility of an umbrella depends on whether it is raining or not (and whether there is a heavy rain or a light rain). Therefore, it would be inappropriate to make an umbrella a prize of a lottery when the nature of the weather is the conditioning event; rather, we must redefine the prize as staying dry, or as getting slightly wet, or as getting very wet—since, as a rule, the utilities associated with these prizes can be assessed without knowing the weather, etc.

(ii) More importantly, both postulates (and especially postulate 3) presuppose that the decision maker has no specific utility or disutility for gambling as such, that is, for the nervous tension and the other psychological experiences directly connected with gambling. In other words, the two postulates assume that the decision maker will take a purely *result-oriented* attitude toward lotteries, and will derive all his utility and disutility from the prizes he may or may not win through these lotteries, rather than from the act of gambling itself.

Clearly, this assumption is seldom, if ever, satisfied in the case of gambling done primarily for entertainment. For example, people who gamble in a casino will usually do this because they are attracted by the

nervous tension associated with gambling; and since the latter may strongly depend on the details of the physical process used to produce the relevant probabilities, they may be far from indifferent to the nature of this physical process. (For example, they may not at all be indifferent between participating in a one-stage lottery and participating in a probabilistically equivalent two-stage lottery.)

On the other hand, it is natural to expect that, in making important policy decisions, responsible decision makers will take a result-oriented attitude toward risk taking. This is probably a reasonably realistic descriptive prediction; and it is certainly an obvious normative rationality requirement as well as a moral requirement: responsible business executives using their shareholders' money, and responsible political leaders acting on behalf of their constituents, are expected to do their utmost to achieve the best possible results, rather than to gratify their own personal desire for nervous tension (or for avoiding nervous tension). Even clearer is the obligation of taking a purely result-oriented attitude in making important moral decisions.

Thus, we can conclude that while postulates 3 and 4 have little application to gambling done for entertainment, they are very basic rationality requirements for all serious policy decisions as well as for personal moral decisions.

II. Expected-Utility Maximization

The main conclusion of Bayesian theory is that a rational decision maker under risk and under uncertainty will act in such a way as to maximize his expected utility; or, equivalently, that he will assess the utility of any lottery to him as being equal to its expected utility (expected-utility theorem). Different authors have used different axioms to derive this theorem, and some of these axioms had somewhat questionable intuitive plausibility.[2] It can be shown, however, that for deriving the theorem, all we need, apart from the two basic probability axioms, are the probabilistic equivalence postulate and the sure-thing principle, both of which represent absolutely compelling rationality requirements for serious policy decisions.

THEOREM 2: *Expected-Utility. If an individual's preferences satisfy postulates* 1, 2, 3, *and* 4, *then he will have a (cardinal) utility function U_i such that assigns, to any lottery L of form* (1), *a utility equal to its expected utility, that is, equal to the quantity*

$$(5) \qquad U_i(L) = \sum_{k=1}^{K} p_k U_i(A_k)$$

where p_k ($k = 1, \ldots, K$) is the probability associated with the conditioning event e_k. More specifically, if L is a risky lottery, then p_k must be interpreted as the objective probability $p_k = Prob(e_k)$ of this event e_k; whereas if L is an uncertain lottery, then p_k must be interpreted as the subjective probability $p_k = Prob_i^*(e_k)$ that the decision maker chooses to assign to this event e_k.[3]

Property (5) is called the expected-utility property, and any utility function U_i possessing this property is called a von Neumann-Morgenstern utility function. For short reference, I will call postulates 1, 2, 3, and 4 the Bayesian rationality postulates.[4]

[2]For example, Leonard Savage's Postulate 4 directly assumes that the decision maker will act on the basis of consistent *qualitative* subjective probabilities (i.e., he will consistently act on the opinion that some events are more likely to occur than not to occur, while other events are more likely not to occur than to occur). In my own view, consistency in the use of qualitative or quantitative subjective probabilities should not be assumed as an axiom, but rather should be inferred from some more basic—and, one may hope, more compelling—axioms. This is the approach taken by F. J. Anscombe and Robert J. Aumann.

[3]Owing to space limitations, the proof has been omitted. But see my working paper.

[4]As M. Hausner has shown, we can obtain a weaker form of the expected-utility theorem without using postulate 2: if an individual's preferences satisfy at least postulates 1, 3, and 4, then his behavior will be equivalent to lexicographically maximizing the expected value of a certain utility vector.

III. An Axiomatic Foundation for Utilitarian Morality

Now I propose to show that the Bayesian rationality postulates, together with a very natural Pareto optimality requirement, logically entail a utilitarian ethic.

According to Theorems 1 and 2, the behavior of a rational individual i ($i = 1, \ldots, n$) would be equivalent to that resulting from the maximization of the expected value of some cardinal utility function U_i expressing his *personal* preferences. In the case of most individuals, these personal preferences will not be completely selfish; but usually they will give greater weight to his own personal interests and to the interests of his family, friends, and other associates, than to the interests of complete strangers.

Yet, there are situations where an individual's behavior will not be guided by his more or less self-centered personal preferences, but rather will be guided by much more impartial and impersonal criteria. We expect that judges and other public officials will be guided in their official capacities by some notions of public interest and of impartial justice; and, more importantly, every individual will have to be guided by certain impartial and impersonal criteria when he is trying to make a moral value judgment. Indeed, by definition, any evaluative judgment based on biased, partial, and personal criteria will not be a moral value judgment at all, but rather will be a mere judgment of personal preference.

I will describe the criteria guiding an individual when he is honestly trying to make an impartial and impersonal moral value judgment as this individual's moral preferences.

In view of Theorems 1 and 2, if the moral preferences of an individual i satisfy certain consistency requirements, then his moral value judgments will be such as if he tried to maximize a special utility function expressing these moral preferences. This utility function will be called his social welfare function W_i.

I now propose to show that a rational individual's social welfare function must be a positive linear combination of all individuals' utility functions. I will assume that society consists of n individuals. Consider the social welfare function of individual j. The following three axioms will be used:

Axiom a: Individual rationality. The personal preferences of *all n* individuals satisfy the four Bayesian rationality postulates.

Axiom b: Rationality of moral preferences. The moral preferences of individual j satisfy the four Bayesian rationality postulates.

Axiom c: Pareto optimality. Suppose that at least *one* of the n individuals personally prefers social situation A over social situation B, and that none of the other individuals personally prefers B over A. Then, individual j will morally prefer A over B.

Axiom a is an obvious rationality requirement. So is Axiom b: it expresses the principle that an individual making a moral value judgment must follow, if possible, even higher standards of rationality than an individual merely pursuing his personal interests. Thus, if rationality requires that each individual should follow the Bayesian rationality postulates in his personal life as postulate 1 asserts, then he must even more persistently follow these rationality postulates when he is making moral value judgments.[5] While Axioms a and b are rationality requirements, Axiom c is a moral principle—but it is surely a rather noncontroversial moral principle.

In view of Theorem 2, Axiom a implies that the personal preferences of each individual i can be represented by a von Neumann-Morgenstern (vN-M) utility function U_i, whereas Axiom b implies that the moral preferences of individual j can be represented by a social welfare function W_j which likewise has the nature of a vN-M utility function. Finally, the three axioms together imply the following theorem.

THEOREM 3: *Linearity of the social wel-*

[5] Axiom b, and, in particular, the assumption that people's moral preferences should satisfy the sure-thing principle, was criticized by Peter Diamond. As I have tried to show in my 1975 paper, his criticism is invalid.

fare function. The social welfare function W_j of individual j must be a real-valued function over all social situations A. and must have the mathematical form

$$W_j(A) = \sum_{i=1}^{n} \alpha_i U_i(A) \quad (6)$$

with α_i strictly positive for $i = 1, \ldots, n$

For the proof, see the author (1955, pp. 313-14).[6]

Note that the proof of Theorem 3 does not assume the possibility of interpersonal utility comparisons. The theorem will remain valid even if such comparisons are not admitted. Of course, if such comparisons are ruled out, then the coefficients α_i will have to be based completely on individual j's personal- and more or less arbitrary— value judgments.

On the other hand, if interpersonal utility comparisons (or at least interpersonal comparisons of utility differences) are admitted, then our three axioms can be supplemented by a fourth axiom:

Axiom d: Equal treatment of all individuals. Individual j's social welfare function W_j will assign equal weights to the utility functions $U_1, \ldots, U_n$ of the n individuals when these utility functions are expressed in equal utility units.

Using this axiom, we can infer that in (6) we must have

$$\alpha_1 = \ldots = \alpha_n \quad (7)$$

IV. The Equiprobability Model for Moral Value Judgments

The axiomatic analysis of Section III has the advantage that it uses only extremely weak philosophical assumptions: it derives utilitarian ethics from two rationality requirements and one very natural moral requirement. However, if we are willing to accept somewhat stronger philosophical commitments, then we can obtain a somewhat stronger form of Theorem 3 (to the effect that the social welfare function must be the arithmetic mean of all individual utilities). What is more important, we can achieve deeper philosophical insights into the nature of moral value judgments.

I have argued that moral value judgments must be based on impartial and impersonal criteria. Now I propose to give a more specific formal definition for this requirement of impartiality and of impersonality.

Suppose individual j expresses a value judgment about the relative merits of one possible social situation A as against another possible social situation B. How do we know whether he expresses a genuine moral value judgment, based on impartial and impersonal considerations? He would certainly satisfy our impartiality and impersonality requirements if he did not know how his choice between A and B would affect him personally and, in particular, if he did not know what his own social position would be in situations A and B. More specifically, let us assume he would think that in either situation he would have the same probability $1/n$ to occupy any one of the n possible social positions—and, indeed, to be put in the place of any one of the n individuals in the society. Then, he would clearly satisfy the impartiality and impersonality requirements to the fullest possible degree. I will call this assumption the equiprobability model of moral value judgments.

Obviously, this equiprobability model cannot be taken literally. When individual j makes a value judgment as to the relative merits of situations A and B, he will often have quite a good idea of the actual social position he would have in each situation; and he will certainly know his own personal identity. Nevertheless, his judgment as to the relative merits of situations A and B will qualify as a genuine moral value judgment as long as he at least makes a serious attempt to *disregard* these morally irrelevant pieces of information in making this judgment.

If we apply Theorem 2 to this equiprob-

[6]In view of Hausner's results, a weaker form of Theorem 3 will remain true even if, in Axioms a and b, we redefine the Bayesian rationality postulates so as to omit postulate 2 (the continuity postulate). In this case both the quantities U_i and the quantity W_j will have to be reinterpreted as lexicographically ordered utility vectors.

ability model, then we obtain the following theorem.

THEOREM 4: *The social welfare function as an arithmetic mean of all individual utilities. Suppose that individual j follows the Bayesian rationality postulates. Then, under the equiprobability model, he will make his moral value judgments in such a way as to maximize the social welfare function:*

$$W_j(A) = \frac{1}{n}\sum_{i=1}^{n} U_i(A) \tag{8}$$

The theorem follows from the fact that under the equiprobability model, j's expected utility will be given by the right-hand member of (8).

Unlike Theorem 3, Theorem 4, as well as the equiprobability model itself, does presuppose the possibility of interpersonal comparisons of utility differences (utility increments).

To sum up, we have found that the Bayesian rationality postulates, together with a Pareto optimality requirement, logically entail utilitarian ethics (Theorem 3)—even if interpersonal utility comparisons are not admitted. But, in actual fact, as I have tried to show elsewhere (see the author, 1977a), there are no valid arguments against such comparisons. Yet, once such comparisons are admitted, the logical connection between Bayesian theory and utilitarian morality becomes even more obvious (Theorem 4).

REFERENCES

F. J. Anscombe and R. J. Aumann, "A Definition of Subjective Probability," *Annals Math. Statist.*, Mar. 1963, *34*, 199–205.

Gerard Debreu, *Theory of Value*, New York 1959.

P. Diamond, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment," *J. Polit. Econ.*, Oct. 1967, *75*, 765–66.

J. C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *J. Polit. Econ.*, Aug. 1955, *63*, 309–21.

———, "Nonlinear Social Welfare Functions," *Theory Decn.*, 1975, *6*, 311–32.

———, (1977a) "Morality and the Theory of Rational Behavior," *Soc. Res.*, Winter 1977, *44*, 623 56.

———, (1977b) "Bayesian Decision Theory," work. paper no. CP-404, Center Res. Manage. Sci., Univ. California-Berkeley 1977.

M. Hausner, "Multidimensional Utilities," in Robert M. Thrall et al., eds., *Decision Processes*, New York 1954, 167–80.

Leonard J. Savage, *The Foundations of Statistics*, New York 1954.

Altruism, Meanness, and Other Potentially Strategic Behaviors

By THOMAS C. SCHELLING*

Biologically, "altruism" poses the puzzle of how a genetically determined behavior that is apparently nonself-serving can favor its own inheritance. Several answers to this puzzle have been found. Equally intriguing is meanness—harmful behavior that serves no evident purpose yet costs something. An illustration was the boy at camp who publicly and viciously hit a "friend" in the mouth while the friend was taking a nap. The act was so threatening that the aggressor became the undisputed leader of his group.

The kinds of behavior I have in mind, of which favors and insults are merely examples, might be called "strategic." A behavior propensity is strategic if it influences others by affecting their expectations. Strategic behavior is ubiquitous in human society and often takes forms that have a paradoxical quality—eliminating options, impairing capabilities, incurring penalties or risks, making expensive demonstrations, destroying possessions, disarming oneself, and other apparently nonself-serving behaviors that are advantageous only in their *anticipation* by others.

To prove that I will not harm you I disarm myself; to keep you from kidnapping my children I have to be poor; to persuade you that I will never bear witness I have to go blind; to keep you from desiring me I have to make myself ugly; to persuade you I shall never retreat I have to be chained to my post. Each is a gratuitous impairment or sacrifice except for the influence it has on your behavior. There have been societies in which the best jobs went to eunuchs because of the confidence their employers had in what they could not do.

I have known since I was a child that bees can sting, and that when they sting they die, and that nevertheless they sting. Unable to explain to a bee that its stinging would merely hurt me but would kill it, I have behaved with great respect toward bees. Scores of bees must have lived, because of my anticipation, for every one that died stinging me.

Meanness and altruism are not only alike, except for sign, but often indistinguishable. Punishing a wrongdoer can be a "public good" motivated as well by meanness as by public spirit. It is even a favor to the victim if it serves as warning, or instills discipline among "victims." That boy who hit a sleeping friend became an instant leader. Maybe they needed a leader.

Strategically there is no difference between behavioral and anatomical constraints. The bee's sting is behavioral; the cactus needle is inert; porcupine quills are both. Cactus is allowed to grow in gardens because it is, though potentially dangerous, very well behaved—it never chases people. Whether it is immobile or merely lethargic doesn't matter. The deterrent effect of being known to defend one's nest to the death will not depend on whether the predator perceives obstinacy, maternal loyalty, or physical incapacity to escape.

I close with a short sample of potentially strategic propensities. For brevity I state each as a declaration of a constraint or incapacity that would be costly if not believed but could be advantageous if believed. But let me clarify my motivation. My interest is not in the possible genetic heritability of any human traits we might identify as strategically self-serving. It is the opposite, to see whether the richness of observable strategic behavior in people can provide hints to the biologists who study other creatures. The fact that people's strategic behavior appears to be conscious, even calculating, does not mean that such behavior has to be conscious and calculating to serve a purpose—rather, to serve strategically,

*Harvard University.

whether or not the creature has a "purpose."

Here is that small sample of strategic declarations; the reader can add dozens more of his own.

I'm harmless, and couldn't hurt you if I wanted to.

I attack anyone who comes near me.

I'll expose us both if you don't do what I ask.

If a predator spots us it will prefer me to you.

I can't run because I'm vulnerable with my back to you.

I'll destroy my property if you approach it.

I will not fight for my property, so you don't have to kill me to get it.

If I do what you ask I'll be punished.

I hurt anyone who fails to help anyone who helps me.

I do not taste good.

You do not taste good.

DISCUSSION

ROGER B. MYERSON, Graduate School of Management, Northwestern Univeristy: John Harsanyi's paper presents us with a remarkable dilemma. He shows that the Bayesian decision theory axioms, if applied to the theory of social decision making, imply that society must use a simple utilitarian criterion to rank possible actions. The Bayesian axioms are widely accepted as reasonable for individual decision theory, and yet many economists are unwilling to limit themselves to utilitarian theories of social choice. To resolve this dilemma we must find at least one of the Bayesian axioms which can be weakened in the context of social decisions. I shall focus my discussion on the sure-thing axiom. I will try to show that a weaker version of it may be more relevant to social decision theory. This will allow maximin or equity-constrained social decision criteria, as well as the utilitarian criteria.

Consider the following example. There are two individuals A and B, two equally likely states of nature θ_1 and θ_2, and there are two actions available in each state. If θ_1 occurs, action α_1 gives utility allocation $(U_A, U_B) = (10,0)$, and action β_1 gives $(U_A, U_B) = (4,4)$. If θ_2 occurs, action α_2 gives utilities $(U_A, U_B) = (0,10)$, and action β_2 gives $(U_A, U_B) = (4,4)$.

Suppose we believe that equity is a fundamental moral concern, so that the best action should be the one which maximizes the individuals' utilities subject to the constraint that their utilities should be equal. Assuming free disposal of utility, this criterion is equivalent to maximizing the minimum utility, or *maximin*. In this example, suppose we judge actions after learning the true state. If θ_1 occurs, then β_1 will be the best action for the maximin criterion, and if θ_2 occurs, then β_2 will be best.

But if we judge action prior to learning the state, then the plan $[\alpha_1, \alpha_2]$ is better than the plan $[\beta_1, \beta_2]$, since the expected utility allocation $(U_A, U_B) = (5,5)$ given by $[\alpha_1, \alpha_2]$ is better than the expected allocation (4,4) given by $[\beta_1, \beta_2]$. So the maximin criterion does not satisfy the sure-thing axiom, because the plan $[\beta_1, \beta_2]$ is sure to look better than the $[\alpha_1, \alpha_2]$ plan after the state is learned, and yet $[\alpha_1, \alpha_2]$ seems better before the state is learned.

This switch in rankings does not necessarily mean that we must disqualify the maximin criterion as a basis for social decision making. It does mean, however, that we must recognize the crucial role which *timing of decisions* may play in determining the outcome. If timing of social decision making can make a difference, then there might be disputes between individuals over when decisions should be made. For the maximin criterion, however, it can be shown that all individuals will always prefer the earlier decision over the later. In the example, the $[\alpha_1, \alpha_2]$ plan selected by maximin in prior analysis does Pareto dominate the $[\beta_1, \beta_2]$ plan selected in posterior analysis.

I wish to propose that the sure-thing axiom could be replaced in social decision theory by a *prior-dominance* axiom, which would say that in any social choice situation, the plan of action which is socially preferred in prior analysis must Pareto dominate the plan of action which would be socially preferred in posterior analysis. Prior-dominance would thus guarantee that if timing does make a difference, then at least it should not be a cause for dispute. The utilitarian criteria will satisfy prior-dominance, because the sure-thing property implies that prior and posterior analysis always select the same plan of action. The maximin criteria also satisfy prior-dominance, when randomized strategies and free disposal of utility are allowed. Furthermore, it can be shown that these are the only two kinds of social preference criteria which can satisfy Pareto optimality and prior-dominance.

JOHN C. HARSANYI, University of California-Berkeley: Many years ago, Thomas Schelling in *The Strategy of Conflict* was the first to point out the possible advantages of making an irrevokable commitment to actions apparently contrary

to our short-run self-interest (by a promise, threat, agreement, self-imposed physical constraint, interruption of a communication channel, etc.). This has been a major contribution to our understanding of strategic behavior, and the concept of a commitment has remained an essential analytical tool in game theory and in related disciplines ever since.

In his note, Schelling argues that the biological analogues of such commitments, viz., the inherited tendencies by animals and by humans to act contrary to their short-run self-interest—which Schelling calls genetically determined "strategic behavior"—may often convey important advantages in the struggle for their survival and for propagating their own genes and therefore may often be favored by natural selection. Many biologists have pointed out that certain forms of altruistic behavior may be advantageous in natural selection. Schelling argues that other forms of strategic behavior (even unselfish malice) may have similar advantages in certain situations.

I feel this point is well taken. I am less happy about Schelling's term "strategic behavior." I think it would be preferable to restrict this term to behavior chosen by conscious strategical calculation. To describe inherited behavioral tendencies I would rather use a term such as "deflective behavior" or "diverting behavior," since the function of such behavior is to deflect or to divert other organisms from the course they would otherwise follow.

No doubt human beings have many nonselfish behavioral tendencies that are biologically and socially useful, or at least used to be useful under precivilized conditions. For instance, most people tend to be helpful, that is, to do minor favors for other people. Likewise, gratitude is a very common human attitude. It is easy to see why these two attitudes are advantageous. Revenge is an attitude of much more dubious usefulness—though under precivilized conditions it may have had its biological value. Anger, a less intensive emotion, probably still is a biologically useful form of behavior in many cases. Human behavior in conflict situations simply cannot be understood without paying attention to these and similar emotions, which often deflect human agents from short-run game-theoretical rationality. One way of fitting such emotions into a game-theoretical model is to assume that they modify the players' utility functions by generating "secondary" utilities and disutilities for gratitude, for revenge, etc. The term secondary utility has been suggested by Reinhard Selten.

In this connection, mention should also be made of the difficulty that most people have in committing certain forms of antisocial behavior. As Konrad Lorenz has pointed out, most animals have a natural inhibition to kill members of their own species, and humans are no exception—though, unfortunately, this natural inhibition seems less effective against "impersonal" killing by modern weapons. By the same token, most people have a natural tendency to tell the truth and, even though they are usually perfectly able to lie, they tend to be rather inefficient liars and often give themselves away by inconsistencies or by involuntary behavioral signs of embarrassment, etc. Once more, it is easy to see why such natural inhibitions are in general socially useful; and they may also be the results of natural selection.

More generally, biology can only benefit by studying possible applications of game-theoretical concepts to ecological relationships. For instance, the notion of a game-theoretical equilibrium point has already made valuable contributions to a study of interspecies equilibrium. Likewise, the concept of strategic (or of deflective) behavior should prove a useful heuristic idea. On the other hand, the study of human behavior can no doubt benefit from judiciously chosen biological analogies—even if discussion of such analogies makes some social scientists unhappy for scientifically regrettable, purely ideological reasons.

ECONOMICS AND BIOLOGY: EVOLUTION, SELECTION, AND THE ECONOMIC PRINCIPLE

The Economy of the Body

By MICHAEL T. GHISELIN*

The process of interrelating disciplines takes many forms. In some cases there is simply an analogizing or a borrowing of jargon—perhaps a mere transfer of metaphor, as when we speak of "the body politic." In other instances one discipline provides useful "tools" for the study of others; for example, physics has given economics and biology units of measurement for the study of energy. Sometimes there is an area of overlap between two fields and a hybrid discipline arises: biochemistry and biophysics are good examples.

Finally we have cases in which two disciplines are really branches of a more general one. The obvious example is zoology and botany forming subdivisions of biology. In the present work I shall take the position that economics and biology do not merely share common interests; they have more than just a few lessons to learn from one another, or an interdisciplinary boundary at which common problems are dealt with. Rather, they constitute a single branch of knowledge.

Just as biology deals with both plant and animal life, there should be recognized a branch of knowledge that deals with economic processes irrespective of whether they are man-made or not, concerning itself with such phenomena as competition that are common to all economies. Thus I propose that we recognize a body of knowledge called *natural economy* (biology) coordinate with *political economy* (economics), together forming a branch of knowledge which we may call *general economy*.

*Department of zoology and Bodega Marine Laboratory, University of California-Berkeley. I thank Brewster Ghiselin, Eric Charnov, and Jack Hirshleifer for advice on the manuscript.

If it be argued that biology is not wholly an economic discipline, I can only answer that it actually is. All the properties of organisms, without exception, are the result of evolution, and the mechanism of evolution, selection, is nothing more than reproductive competition between members of the same species. Competition, of course, is as fundamental an economic phenomenon as can be imagined.

Of course, there are differences between natural and political economies. There is no analogue of sex in economics. Biology has nothing strictly equivalent to money. But subdisciplines often have phenomena peculiar to themselves; animals do not photosynthesize sugars, and plants lack nerves.

The main reason for recognizing the unity of biology and economics is the advantage gained when generalizations on a high level can be applied across the board. Furthermore, the greater diversity of systems thus made available allows better comparative study. Some phenomena are best studied in the natural economy, others in the political one. By way of illustration, I have selected the example of how the principle of the division of labor can be applied to anatomy. It happens that natural selection has produced numerous and instructive examples of this phenomenon.

It seems odd that biologists and economists alike have paid very little attention to the division of labor. Seeming to be too obvious to require explanation, it has been accepted as a mere brute fact, while its functional significance has been virtually ignored. Although labor is sometimes divided, sometimes combined, there are as yet no adequate explanations why.

Economists have written surprisingly lit-

tle on the division of labor. Elementary text books do not treat it in detail, and only a few recent papers have dealt with the topic (George Stigler; Hendrick Houthakker). Bernard Haley has reviewed the literature, and provides a few more references. This neglect is all the more curious, because the division of labor was one of the pillars of Adam Smith's system.

Turning to *The Wealth of Nations* itself, we find that it contains an exceedingly simpleminded discussion. It would seem that although he recognized that the division of labor increases industrial output, Smith took little interest in just how this happens. In fact, he gives only three advantages to the division of labor, none of them particularly profound. *First*, it increases dexterity through practice. This is probably a fairly important advantage. *Second*, it saves time, because it avoids shifting from one task to another. While this may be significant in some cases, shifting tasks does not consume a great deal of time. *Third*, concentration upon a single task makes conditions more favorable for technological innovation. It seems dubious whether this has much significance. Smith mentions no further advantages to the division of labor.

Perhaps we should go straight to biology to see how little Smith's explanations apply. A good example is the division of labor among teeth in the human mouth. We have incisors for cutting, canines and premolars for shearing, and molars for grinding. The reason for this differentiation is obvious enough. A different shape of tooth is optimal for each kind of chewing, and a tooth cannot change its shape. Smith's explanations doubtless have some validity, but I believe that the kind of explanation just proposed for teeth will prove to have wider applicability in both natural and political economy. But let us consider a little more economics before going on with biology.

Imagine, if you will, a man who tried to function as both a plumber and a carpenter. He would have to invest time in learning both trades, and money in purchasing the tools for both. Since he could not ply both trades simultaneously, the return on his investments in training and tools would be low. Hence, given a sufficient extent of the market, it would be most profitable for him to specialize in one or the other activity. Yet is is conceivable that conditions exist in which such division of labor is not advantageous—for example, when one training or one set of tools qualifies the artisan for more than one profession, or when both activities can be carried out simultaneously. The former case was mentioned in passing by Smith; his example is teaching combined with research. Time spent reading learned books and monographs can be applied to both lectures and scholarship. As to the second case, we have a few activities that can be carried out simultaneously without their interfering with each other; studying and baby-sitting are a good example.

In the natural economy, we should find division of labor wherever generalists would have to possess a multiplicity of "tool kits." Such division of labor between species is evident in many cases of "adaptive radiation" among animals. The gastropods (snails and slugs) for example have diversified enormously (they form the second largest class in the animal kingdom). Much of this diversity is due to the fact that the feeding apparatus can be modified in a remarkable variety of ways, allowing each species to have an instrument highly effective for utilizing a particular kind of food. Here the division of labor allows a lot of gastropod niches (professions) to be exploited with profit. This is an example of the sort of "competitive division of labor" which occurs between firms, not the "cooperative division of labor" like that within firms (see the author, 1974a). Ecologists have devoted much effort to explaining "species diversity" but for the most part at least have not invoked such organismal reasoning. Virtually all of their theory focuses upon higher levels of integration, which seems to me a mistake (see below). Likewise the notion of "extent of the market" has scarcely been considered in dealing with such problems and the term, to my knowledge, has never been used in biology. I have attempted to show how this

kind of reasoning is useful in dealing with species diversity and related topics, such as the size of insect societies (see the author, 1974a).

Whether labor will be divided or combined in the body depends on whether there is interference between the different tasks being carried out by the organs in question. An interesting comparison is to be made between mammals and birds. In the mammalian mouth, labor is combined. This organ serves both as an instrument of grasping and subduing prey, and as a mechanism for its mechanical breakdown. Birds, for reasons we need not go into, have on the other hand lost their teeth, and have a separate organ, the gizzard, that breaks food down into small particles. One may conjecture that the need for this extra organ wastes a certain amount of material and energy.

In evolutionary series there seems to be a definite trend toward the specialization and effective spatial organization of organ systems. Consider, if you will, the gut of a very primitive organism, a flatworm. It is set up so that only one alimentary function can take place at one time. The system is a simple cavity with a single opening, usually with a mechanism for grasping food. The prey has to be taken up, then broken down, then absorbed, and finally waste products voided all in the same place; it is like a work room. Now compare a more advanced organism, a snail. The mouth is provided with a specialized organ (the radula) for the uptake of food. It connects via an oesophagus to an organ, the crop, which serves to store food prior to further processing. Next comes a gizzard, which breaks down the food mechanically. Afterward is a stomach, which completes the digestion on a chemical level, and then a digestive gland where the nutriment is absorbed. An intestine and anus provide for the disposal of waste. Here we see that the functions are discharged in a series, as on an assembly line.

It seems to me that one of the major reasons for the division of labor in the body is that it prevents mutually incompatible activities from interfering with one another (see Houthakker). In my doctoral research (1966) I studied the reproductive anatomy of a group of gastropods which had become hermaphroditic. Primitively, a single duct had to do multiple service. It had to transport outgoing sperm, incoming sperm, and eggs. It is easy to imagine how these various products could get in one another's way. Not surprisingly, there evolved separate ducts for each of the three genital products.

Labor is often combined where functions are intermittent and need not be carried out simultaneously. Thus in many offices typists or file clerks are given the added responsibility of answering the telephone. My favorite biological example of the combination of labor is a flatworm which uses his and her penis to subdue prey. In many animals, including male human beings, the same duct serves for the transport of both genital and excretory products. Thus where functions do not interfere with one another there is no advantage to a division of labor, and its combination is more economical. Comparative anatomy teaches us that functions have been combined and redivided in the course of evolutionary history. Thus in the annelid superphylum (Trochozoa) the genital and excretory tracts were originally separate. When a body cavity evolved, it was possible for the products of both systems to exit via a single duct and the two were united. At this stage fertilization was a simple process: both sperm and eggs were released into the water where they united. Later, fertilization became internal and protective layers were secreted around the eggs. This rendered the common ducts ineffectual and secondary specialization resulted. This phenomenon in not unique; the anus, for example, has been secondarily lost several times. Similar conclusions may be drawn from the union and separation of the sexes (reviewed in the author, 1974a). Originally hermaphroditism was thought to be primitive, but we now know that separate sexes came first, at least in animals. Having both sexes simultaneously is advantageous under special circumstances, namely when it is hard to find a mate. Interestingly enough the sexes are occasionally divided temporally, with an animal changing from a male into a female, or vice

versa. This phenomenon resembles the change in habitat structure and way of life in invertebrate life histories. In some cases changing tasks may be advantageous. Such occurrences have two major implications. First, it is not just the extent of the market that determines the division of labor, but also the efficacy of production. Second, division of labor is not, as early economists and biologists believed, a unidirectional trend, nor an advanced condition.

Within the phylum Arthropoda, there is an interesting evolutionary trend, beginning with animals having many similar appendages, and going toward animals with fewer, differentiated ones. Probably the reason why the series is not differentiated from the beginning is that, historically, the arrangement with many similar parts originated as a means of producing a large organism. There is some advantage to having a lot of similar appendages. They can substitute for one another and discharge their functions locally. On the other hand, higher arthropods tend to have their limbs differentiated into sense organs, mouth parts, walking or swimming limbs, respiratory organs, and genitalia.

The foregoing are only a few rather elementary examples of how the principle of the division of labor may be applied to organismal biology. A sketchy approach has been necessary because biologists have given it virtually no consideration whatsoever. Aristotle and, early in the last century, the Belgian physiologist Henri Milne Edwards gave it passing mention, but I know of no papers dealing with it on the organismal level. On the other hand there does exist a modest literature on division of labor among insect societies. As in economics, very little effort has been made to provide functional explanations for the observed phenomena. The field remains descriptive, rather than nomothetic. As I have reviewed this literature at length elsewhere (1974a), I shall not say much about it here. It is worth noting, however, that labor is divided both between and within castes, as when some worker bees fetch water while others go after nectar. However, the workers within a hive each seem to engage in a variety of tasks. Furthermore, the younger bees work in the hive, but switch to foraging when they get older, probably because life outside is dangerous and aged individuals are more expendable. Such "temporal division of labor," already alluded to, may be more widespread than has been realized. Certainly it makes sense to engage in different activities as one gets older, and as the economy changes (as with the seasons).

It is interesting to inquire into the reasons for this curious neglect of so important a subject as the division of labor. The situation appears all the more odd when we realize that Darwin himself attributed his discovery of natural selection to reading a work by the economist Malthus. It should have been obvious that there were further opportunities, but these were not followed up. Biologists have borrowed little more than analogies and jargon from economics. Only recently have a few ecologists and evolutionists applied economic models to their subject. But such actual transfer between disciplines as has occurred has been "retail" not "wholesale" in scale. Important discoveries within biology have indeed been made, but these seem to have been more the result of "convergent evolution" than "hybridization." My book *The Economy of Nature and the Evolution of Sex* was, on the contrary, a deliberate and systematic exercise in interdisciplinary transfer and synthesis. Unfortunately, certain persons have preferred to interpret a work on epistemology and metaphysics as if it were a biological monograph. (It is as if they were to consider this paper "nothing more" than a discussion of the division of labor, which it definitely is not.)

An excessive academic division of labor with too little communication between fields provides only a partial explanation. I think the fundamental reason why biologists so long failed to think like economists is the manner in which they happened to conceptualize natural selection. There are many ways of thinking about evolution; some good, some bad, some indifferent. One way to reason about natural selection is in terms of Herbert Spencer's "survival of the fittest." Now it is certainly important for animals and

plants to survive. The phenomenon has an economic analogue in avoidance of bankruptcy. However focusing attention on this rather peripheral issue causes one to lose sight of a more fundamental one. What really matters in natural selection is reproductive output, which is analogous to profits in economics. Organisms should be compared to factories, not battleships. Once one realizes, as Darwin did, and as modern biologists are coming to realize too, that selection is reproductive competition between components of the same species, the close connections between the two fields become much more obvious.

Subtler conceptual reasons, discussed at length in my book (1967a) may also be invoked. These, indeed, are the basic point of my work, and deserve special consideration. Biologists and economists alike deal with groups and units that exist at levels of organization: cells, organs, organisms and species on the one hand; employees, firms and markets on the other. Biologists have been confused about the nature of the units and the significance of the levels. Thus they have tried to explain many phenomena in terms of the adaptation of species, when they should have been explaining them in terms of organisms. In economics firms are analogous to species, employees to organisms (see the author, 1974b). The philosophical doctrine that I call "radical individualism" should be equally enlightening to both fields.

This all leads up to what is perhaps my most important point. All sciences encounter difficulties with what is called "metaphysics." The Cambridge economist Joan Robinson has analyzed her field from a metaphysical point of view and concludes that her discipline is a branch of ethics trying to become a science. If economists disassociate themselves from political science, and align themselves with biology, not only will this metaphysical problem be resolved, but much important synthesis can be looked forward to. Even so, both fields will still have to face up to another metaphysical problem: teleology. In my doctoral research I presupposed that changes would result in functional improvement. To a considerable extent this was true. As I have mentioned, division of labor frequently led to separate ducts. Then I came to a bizarre group called the sacoglossans. To be sure, the reproductive system worked, but the arrangements were absurd. For instance, some had given up copulation via the usual orifice, and taken to hypodermic impregnation, with the sperm reaching the eggs via the circulatory system. The whole contraption looked like it had been designed by an idiot. It was: natural selection has an IQ of zero. I was encountering an instance of maladaptation, or what Ernst Haeckel called "dysteleology." It is not uncommon in the complex reproductive systems of flatworms. Nor is it in the least unusual in man. The ducts which drain the human bladder are in just the right place—for a quadruped but not for a biped. Hence I end on a cautionary note. Beware of the Panglossian notion that this is the best of all possible worlds. Realize the limitations of economic rationalism, and watch out for the pitfalls of optimization techniques.

REFERENCES

Michael T. Ghiselin, "Reproductive Function and the Phylogeny of Opisthobranch Gastropods," *Malacologia*, June 1966, *3*, 327-78.

——, (1974a) *The Economy of Nature and the Evolution of Sex*, Berkeley 1974.

——, (1974b) "A Radical Solution to the Species Problem," *Syst. Zool.*, Dec. 1974, *23*, 536-44.

B. F. Haley, "Specialization and Exchange," in David L. Sills, ed., *International Encyclopedia Social Sciences*, Vol. 15, New York 1968, 111-15.

H. Houthakker, "Economics and Biology: Specialization and Speciation," *Kyklos*, 1956, *9*, 181-89.

Joan Robinson, *Economic Philosophy*, Chicago 1968.

Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, London 1776.

G. Stigler, "The Division of Labor is Limited by the Extent of the Market," *J. Polit. Econ.*, June 1951, *59*, 185-93.

Competition, Cooperation, and Conflict in Economics and Biology

By J. HIRSHLEIFER*

From one point of view, the various social sciences devoted to the study of man, economics among them, constitute but a subdivision of the all-encompassing field of sociobiology (see Edward O. Wilson, 1977). This idea need not be strange to economists. Adam Smith rested what he regarded as the foundation of human economy, the division of labor, upon a (supposedly unique) biological instinct of mankind to truck, barter, and exchange. And Alfred Marshall declared that economics is a branch of biology. But a rather more exciting interpretation of the relation between the two fields of study has been proposed here by Michael Ghiselin: he argues that biology should be regarded as *natural economy*, while much if not all of the subject matter of the conventional social sciences comprises in essence the field of *political economy*—both being subdivisions of universal *general economy*.

I. Natural Economy vs. Political Economy

In traditional political philosophy, or legendary political history, the step from natural economy to political economy was taken only by man in the form of the social contract of Rousseau or Hobbes. But one thing we have learned from comparative sociobiology is that there is no such sharp discontinuity in social organization, just as there is no sharp discontinuity in physical form between man and other branches of life. Within a social group, *law* emerges when "moralistic aggression" (see Robert Trivers) by third-party intervenors serves to control internal strife. Parental regulation of intrafamily conflict is an obvious and widespread example. *Government* may be said to exist when, in groupings larger than a single family, control tasks are performed by specialists (in the biological realm, generally by dominant animals). The immunities from invasions thus created prefigure the human institution of property (see Melvin Fredlund).

The advantages of these political economy institutions are evident. Law and government deter or limit the internal fighting that would be disfunctional for the group as a whole. Individuals need not divert effort to continual patrolling and monitoring. Also, with property recognized, *exchange* becomes a possibility—and, ultimately, the more sophisticated dealings in deferred reciprocations that constitute the essence of *contract*.

Yet the institutions of political economy can never be so perfect as to entirely displace, even in human societies, the underlying realities of natural economy. Every living organism remains to some degree in a Hobbesian state of nature. Most importantly for mankind, intercourse among nations lies outside the scope of effective law. Even under law and government, the rational self-interested individual will strike a balance between lawful and unlawful means of acquiring resources—between production and exchange on the one hand and theft, fraud, and extortion on the other. For that matter a perfectly law-abiding individual (if there is any such) could not have such confidence in third-party enforcement as to entirely forego personal vigilance and self-defense. And setting aside *violation* of law, the structure of the law itself will necessarily have greater or lesser imperfections. It is not always practicable to define rights to property in such a way as to ban socially wasteful activities designed to capture benefits while imposing costs on others (what we call externalities), or to foreclose efforts aimed at influencing government or revising the law so as to redefine rights in one's favor. This latter activity is of course

*University of California-Los Angeles.

the stuff of redistributive politics. In short, while the intellectual division of labor whereby biologists concentrated on natural economy and economists on an idealized political economy is an entirely understandable one, in the actual world the separations are by no means clean-cut.

As categories in general or universal economy, fundamental concepts like scarcity, competition, equilibrium, and specialization play similar roles in biological and economic systems. Analogs are suggested by terminological pairs like species/industry, mutation/innovation, mutualism/exchange, and evolution/progress. Regarded more systematically, the isomorphism between formalizations of natural economy and political economy—between sociobiology and economics—involves the intertwining of two levels of analysis. On the first level, the acting units or entities choose strategies or develop techniques that promote success in the struggle or *competition* for advantage in given environments. The economist usually calls this process "optimizing," the biologist, "adapting." The equations involved represent constrained maximization. The second, higher level of analysis examines the aggregate result, the social consequences of the interactions among the striving entities. Here we have equations of equilibrium, or of paths of change toward solution states. The solutions on the two levels are of course interdependent. The pursuit of advantage on the part of acting units takes place subject to opportunities and constraints that emerge from the social context, while the social configuration itself depends upon the strategies employed by the advantage-seeking entities.

II. Competition in Relation to Cooperation and Conflict

Competition is the all-pervasive law of natural economy interactions. The source of competition is, of course, the limited resource base of the globe in the face of the universal Malthusian tendency to multiply. By natural selection the biosphere has come to be filled by life forms successful at multiplying and pressing upon one another for command over resources. This teeming of life is therefore both cause and consequence of biological competition.

Biologists have found it useful to distinguish between "scramble" and "interference" as competitive strategies. *Scramble* competitors ignore one another, interacting only through depletion of resources. The winning organisms are those most efficient at extracting energy and other inputs from the external environment. *Interference* strategists, in contrast, gain and maintain control over resources by fighting off or reducing the efficiency of rivals. By a further extension, an interference strategist may go so far as to become a *predator*—for whom the competitor organisms become part of the resource field.

From these biological categories it might be inferred that competition must be wasteful and antisocial (Pareto inefficient). And yet the economist views competition as essentially a harmonizing force (the "Invisible Hand"). How can this be? First, under (idealized) institutions of political economy only scramble competition would be permitted. A businessman cannot blow up his rival's shop (interference), or stock his own store by raiding the other's inventory (predation). (Note that "predatory competition" in the antitrust sense is an inaccurate biological metaphor.) And furthermore, again ideally speaking, the assignment of property rights would be such as to rule out "Pareto-relevant externalities." If this condition were not met, then, as in the classical fishing model of H. Scott Gordon, even the more innocuous scramble competition has each fisherman impairing the extractive efficiency of his rivals. There would thus be a net social loss from competitive (as opposed to monopolized) fishing effort.

The second essential is that competition for the economist is always a *three-sided* interaction: vying *against* a rival or rivals, but *for* the opportunity to engage in mutually advantageous exchange with a third party. Biological competition *may* also be three-sided, as when males vie to

mate with females, or flowers produce fragrances to attract pollinating insects. More commonly, biological competition is two-sided striving, which inevitably becomes socially wasteful. A human example would be duels for survival, as between gunfighters or nations, conducted of course under rules of natural economy. The Invisible Hand thus requires a severely constrained form of competition: vying to engage in exchange with third parties, and doing so only by offering better terms under an ideal system of property and law.

III. Cooperation, Conflict, and Their Limits

If competition is the basic law of life, how is it that organisms sometimes confer benefits on others? In the political economy cooperation mainly takes the form of exchange for mutual gain. How can the analog of exchange, mutualism, be viable in the realm of natural economy without any system of law to guarantee repayments? And how can one-way transfers, unilateral gifts, be explained in either the natural or the political economy? Or, where *conflict* exists, how and why is it that the battle is often limited rather than all out (see Konrad Lorenz)?

The economist approaches problems of optimal action by distinguishing between *preferences* and *opportunities*. Traditionally, he takes preferences as arbitrary brute facts. An individual might inexplicably have an innate sympathy for some of his fellow men and, perhaps, an antipathy to others. Yet to a good approximation, mainstream economics has argued, people can be assumed to be merely neutral or self-interested (economic man). In our intellectual tradition, overwhelming emphasis has consequently been placed upon the determining importance of *opportunities*: the mutual advantages that entirely self-interested persons can gain through social cooperation, to wit, increased production through specialization and the division of labor and improved allocations of products through exchange.

But, social biology tells us, preferences (and, in particular, innate benevolence or malevolence) are *not* arbitrary but are themselves mainly adaptive (see Darwin on social instincts in *The Descent of Man*, chs. 3–4). The fundamental brute fact is *the selfishness of the gene* (see Richard Dawkins). Man himself, full of love and hate and sheer cussedness, ill fits the model of "economic man"—but the gene is an "economic gene." It has been selected to survive on the basis of successful selfishness. However, depending upon *opportunities*, the interests of the gene may sometimes be served if the organism housing it is programmed to help or to hurt other organisms.

Biologists have examined the problem of social cooperation (see especially Mary Jane West Eberhard, Trivers, Wilson 1975) under the heading of "altruism," followed by some economists (Gary Becker; Mordecai Kurz). But altruism is a term with unwanted and rather misleading connotations, and I will instead speak here simply of *the determinants of helping*. The patterns of helping are grouped by biologists into three main categories: those associated with *kinship*; those merely *incidental* to selfish behavior; and those involved in *reciprocal* interactions.

In the most straightforward kinship case the basic helping rule (see W. D. Hamilton) says that evolutionary selection impels a donor D to aid a recipient R if $c_D < r_{DR} b_R$, if the *cost-benefit ratio* of the action (c_D/b_R) is less than the *degree of relatedness* r_{DR} between the pair. Both benefit b and cost c are measured here in terms of increments to reproductive survival—the ultimate payoff that biologists call "fitness" (W). Relatedness constitutes the innate genetic preference element: the helping gene in organism D values itself equally with any identical copy of itself, and r_{DR} measures the chance of organism R having an identical copy. Thus, other things equal, a gene for kinship helping instructs a man to give his life to save two siblings, four half-sibs, eight cousins, etc.

Of course other things are not in general equal; the c/b ratio expresses the individual's "terms of trade" or *opportunities* for helping. While r is a constant feature of

the interaction between two organisms, the c/b ratio may be rather volatile. Nevertheless, it is sometimes possible to characterize classes of interactions where aid is cheap (low c) or benefit is big (high b), and observation generally confirms the prediction that helping evolves in such situations (see West Eberhard). One example relevant for humans: because offspring generally need help more urgently, and parents are in a position to give it, from cost-benefit considerations we would expect to see parents aiding children more than children aiding parents, even though relatedness r is the same both ways. Some biologists have argued that kinship helping can scarcely be important beyond the immediate family, as relatedness r falls off very rapidly toward zero. But West Eberhard points out that one individual might sometimes be able to influence fitness of a great many others, so that increases in numbers affected may offset decrease in average r. In human endeavors this perhaps explains the grueling hours and intense devotion often observed of leaders in war, politics, or even business.

The interpersonal-fitness opportunity set might be more or less *competitive* (unrewarding to mutual aid). In the limiting case of severe competition, the environment can only support a fixed number of organisms. Then helping some necessarily means hurting others. In this situation each organism optimizes by transferring fitness, subject to diminishing returns, from individuals less closely related to those more closely related to him (including himself in the calculation). The xenophobic implications are somewhat mitigated, however, to the extent that closeness of competition tends to be positively correlated with relatedness.

Looking at this another way, we can suppose that donor D is maximizing his "inclusive" fitness $W_D^* = W_D + r_{DR}W_R$, subject to a given resource endowment, and a "production function" showing the alternative combinations of W_D and W_R he can bring about (Eric Charnov, personal communication). The usual constrained maximization procedure leads immediately to Hamilton's helping rule. However, as the increment to recipient R's "fitness income" will generally lead R to *react* (depending on *his* preferences and opportunities) in such a way as to help or hurt the donor or third parties, donor D should take these reactive *income effects* into account. The "rotten-kid theorem" (see Becker, and the comment by the author) provides an instructive instance.

Reactive interactions are also relevant for *incidental* and, especially, *reciprocal* helping. Ruling out the kinship element by letting $r_{DR} = 0$, the helping rule reduces simply to $c_D < 0$. But c_D can be decomposed into a primary cost term c_D^0 and a reactive cost term c_D'. Incidental helping is associated with negative c_D^0 (the helping act has a primary selfish net benefit). Reciprocal helping is associated with negative c_D' (the repercussion upon D of R's reaction to aid is favorable). Evidently, in each case it is the *sum* of the decomposed cost elements that determines whether the helping rule is met. More generally, all three elements—kinship (positive r_{DR}), incidental (negative c_D^0), and reciprocal (negative c_D')— may be involved where helping is observed.

Reciprocation brings us to the interactions mainly studied by economists. The basic rule of reciprocal interactions is "help your helper!" But two classes of reciprocations can be distinguished. First, it may be that recipient R is an individual who would *unconditionally* help you in return. His reactive income effect for helping you is positive (and, of course, sufficiently large). This requires that he be, on the margin, a kinship or incidental helper to you. But where R is not an unconditional helper, there may still be a *potential* gain from mutual aid (i.e., from exchange). If R were to react positively, D would help so that both gain, but what if, *ex post*, R is not motivated to reciprocate?

The shift from natural economy to political economy solves this Prisoner's Dilemma problem via *third-party enforcement of contract*. But political economy institutions are inevitably imperfect. Observed human improvisations and substitutes for enforcement of contract are often paralleled by evolutionary emergences in

the nonhuman sphere: 1) *Familial cooperation:* Relatedness may assure a degree of reciprocation. In less advanced human societies, business associations (firms) extending beyond a single family are almost unknown. 2) *Merger:* To the extent that a group shares a common fate, helping others becomes self-help. Merger can be of the *complementation* or *supplementation* types, involving specialization in the former case (for example, males and females) and returns to scale in the latter (as when a group's fighting prowess depends upon its size). 3) *Repeat business:* Where reputation or "brand name" can be acquired, the prospect of future association reduces the gain from cheating. 4) *Conditional commitment:* As in military deterrence theory, guaranteeing to repay (even though irrational *ex post*) good-for-good, or evil-for-evil may limit conflict or enforce cooperation. Mechanisms of conditional commitment include uncontrollable emotions such as rage (which perhaps explains survival of something often thought to be totally disfunctional). 5) *Ethics and indoctrination:* Uniquely, perhaps, man has evolved a capacity for cultural indoctrination toward cooperative behavior (Wilson 1975, ch. 27).

Under ideal political economy institutions, the Coase Theorem (see R. H. Coase) guarantees Pareto-efficient solutions. Without these institutions there are make-shift solutions as just described, but the inability in natural economies to fully control force and fraud limits the prospects for efficiency. (But see H.E. Frech III.) An alternative process rewarding cooperation is *group selection*: associations of cooperating organisms may win out in social competition against more disunited groups. However, group selection for helping is only rarely effective in the biological realm. What economists would call "free riding" inevitably selects for selfishness *within* the group (Wilson 1975, ch. 5; Dawkins, ch. 6). On the other hand, free riding also serves to limit malevolent activity—since, if *A* expends fitness to hurt *B*, some of the benefit may be reaped by *C* (see Geoffrey Blainey, ch. 4; Dawkins, ch. 5; Gordon Tullock). Group selection may, however, explain the extraordinary sociality of ants and bees, or the observed tendency toward reduced virulence of disease bacteria. Conceivably, group selection under primitive conditions may have led to the evolution of instincts favoring in-group cooperation and out-group hostility among humans (see Richard Alexander).

REFERENCES

R. D. Alexander, "The Search for a General Theory of Behavior," *Behavioral Sci.*, 1975, *20*, 77–100.

G. S. Becker, "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," *J. Econ. Lit.*, Sept. 1976, *14*, 817–26.

Geoffrey Blainey, *The Causes of War*, New York 1973.

R. H. Coase, "The Problem of Social Cost," *J. Law Econ.*, Oct. 1960, *3*, 1–44.

Richard Dawkins, *The Selfish Gene*, New York 1976.

H. E. Frech III, "Biological Externalities and Evolution: A Comment," *J. Theort. Biology*, 1973, *39*, 669 72.

M. C. Fredlund, "Wolves, Chimps, and Demsetz," *Econ. Inquiry*, June 1976, *14*, 279-90.

M. T. Ghiselin, "The Economy of the Body," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 233–37.

H. S. Gordon, "The Economic Theory of a Common-property Resource: The Fishery," *J. Polit. Econ.*, Apr. 1954, *62*, 124–42.

W. D. Hamilton, "The Genetical Evolution of Social Behavior," *J. Theoret. Biology*, 1964, *7*, 1–16.

J. Hirshleifer, "Shakespeare *vs.* Becker on Altruism: The Importance of Having the Last Word," *J. Econ. Lit.*, June 1977, *15*, 500–02.

M. Kurz, "Altruistic Equilibrium," in Bela Balassa and Richard Nelson, eds., *Eco-*

nomic Progress, Private Values, and Policy: Essays in Honor of William Fellner, New York 1977.

Konrad Lorenz, *On Aggression*, New York 1966.

R. L. Trivers, "The Evolution of Reciprocal Altruism," *Quart. Rev. Biology*, Mar. 1971, *46*, 35–57.

G. Tullock, "Altruism, Malice, and Public Goods," *J. Social Theor. Structures*, forthcoming.

M. J. West Eberhard, "The Evolution of Social Behavior by Kin Selection," *Quart. Rev. Biology*, Mar. 1975, *50*, 1–33.

Edward O. Wilson, *Sociobiology*, Cambridge, Mass. 1975.

———, "Biology and the Social Sciences, *Daedalus*, Fall 1977, *106*, 127–40.

DISCUSSION

R. H. COASE, University of Chicago Law School: I would not be participating in this session unless I thought that developments in biology were important for economists. But my reasons for believing this are very different from those which lead Michael Ghiselin to his enthusiastic endorsement of the co-mingling of economics and biology. Ghiselin argues that economics and biology "constitute a single branch of knowledge" and their treatment as such will enable "generalizations on a high level" to be "applied across the board." Whether biologists should study economics, I do not know. What matters for me is whether economists should study biology. If what we would gain are simply generalizations equally applicable to biology and economics, I see little advantage in it. Our theories are, for most purposes, already too general. While we would doubtless learn something from a theory which is equally instructive in explaining the division of labor in the oil industry, among the teeth in the human mouth, and in a community of ants, our pressing need is to explain the division of labor in the oil industry. It is not general theories which we lack, but theories which explain the working of our actual economic system. Ghiselin rightly points out that our analysis of the division of labor is primitive. The reason is not a lack of knowledge of biology but of economics. We likewise have a primitive analytical system to handle the firm, the market, the process of contracting and property rights—all vital elements in the working of our economic system.

In seeking to improve our treatment of such subjects, biological analogies can be very misleading. The firm, the market, the legal system are all social institutions and are the result of purposeful human activity. Ghiselin explains that natural selection has an IQ of zero. The IQ of businessmen and politicians may not be high, but it is not zero. Natural selection produces its results by trial and error over long periods of time. Economic systems, such as the structure of an industry, may be transformed within a single generation.

This may sound as though I thought that biology was not likely to have, and should not have, much influence on economics. This is not so. But the place where its immediate influence will probably be felt is not in the analysis of competition, but in utility theory. Modern utility theory, which analyzes the effect of human preferences on economic behavior, largely regards man as a rational utility maximizer. It tells us little about the purposes which impel people to action. Recently, attempts have been made to relate preferences to certain basic needs. But it seems unlikely that we can make much progress without a comprehensive view of man's nature on the basis of which such a theory can be developed. It is such a view which is found in sociobiology.

Human nature is here seen as the product of evolution over a long period and is genetically determined. The structure of human nature includes learning rules, and what is particularly important in economics, the way in which we translate experience into expectations. Human nature is what it is because it contributed to human survival in the conditions in which human beings developed, not, as is sometimes supposed, because it contributed to human happiness, although what may be termed altruistic behavior may have survival value. This approach downplays the rational element in human behavior. As Ghiselin has said, "If fools are more prolific than wise men, then to that degree folly will be favored by selection" (see his *The Economy of Nature and the Evolution of Sex*, p. 263). This approach is similar to that of Adam Smith. Jacob Viner noted in his article in *1776–1926 Lectures to Commemorate Publication of "The Wealth of Nations,"* the lowly role which reason plays in Adam Smith's thought: "Under normal circumstances, the instincts make no mistake. It is reason which is fallible." Thus, as Adam Smith said in 1759 when discussing self-preservation and the propaga-

tion of the species (both essential to survival), they are so important that they have not been "entrusted to the slow and uncertain determinations of our reason" but to "original and immediate instincts." Furthermore, as Ghiselin has noted, Adam Smith was very much aware of the adaptive or optimizing aspect of human psychological characteristics. In the refining of economic theory that has occurred since Adam Smith's time, this idea has been lost (along with other good things in his work). One result of sociobiology may be to lead us back to our founding father. In any case, it should have a profound and pervasive effect on economic analysis. The danger is that dabbling in sociobiology may prove to be more attractive to many economists than the use of sociobiological findings to improve our economics.

The Economics of Special Interest Politics: The Case of the Tariff

By WILLIAM A. BROCK AND STEPHEN P. MAGEE*

This paper investigates the interaction of politicians and special interest lobbies. The goal is to model the equilibrium levels of redistributive policies such as taxes, subsidies, tariffs, and regulatory decisions. The model is applied throughout to the setting of tariffs. By contributing to political campaigns, special interest lobbies generate economic returns: higher tariffs generate revenue for import-competing industries through higher product prices, higher milk support prices raise dairy farm income, and so forth. The lobby invests to maximize these net economic returns. In evaluating their chances of election, politicians weigh the favorable effects of special interest money against the unfavorable association with the lobby and the social cost of the redistributive policy. Section I explores the lobby's problem: it describes a formalization of Mancur Olson's discussion of the voluntary provision of public goods and optimal contributions by lobbies to political campaigns. Section II describes determination of the equilibrium tariff positions of each politician in a political campaign.

I. The Lobby

Modeling the participation of economic agents in political activities is similar to modeling consumer behavior. In both cases, there are important subjective and nonobservable elements. In demand theory, we deal with these problems by imposing weak assumptions on the nonobservable constructs (for example, monotonicity and quasi concavity on the utility functions) to get general results or use stronger assumptions (separability) to derive richer implications. Our formulation of the lobbying problem follows the second of these two approaches.

The first problem in tariff analysis is to identify the gainers and the losers. If a tariff is increased from 0 to T, the stakes of each participant are defined as the difference which the tariff makes in the present discounted values of their income streams. There are two polar models in international trade theory predicting the redistributive effects of tariffs. The Stolper-Samuelson model suggests that the intensive factor in import-competing activities gains throughout the whole economy from a tariff (say, labor), while the other factor in the economy loses (say, owners of capital). The model of James Cairnes assumes that factors are completely immobile and hence sector specific, so that capital and labor in import-competing industries gain from a tariff while capital and labor in export industries are worse off. A study of the lobbying behavior of *U.S.* industries during hearings for the U.S. Trade Act of 1973 indicated that American capital and labor are quite sector specific (see Magee). Thus, contrary to Stolper-Samuelson, tariff-induced increases in the price of an industry's output benefit both capital and labor in the industry.

Consider the potential gainers from a tariff in a specific industry. What assumptions about the behavior of the n gainers guarantee Olson-type lobbying behavior in a mathematically consistent noncooperative game theoretic framework (the behavior of the losers is symmetric)? It is

*Professor of economics, universities of Chicago and Wisconsin, and professor of finance, University of Texas-Austin, respectively. Mathematical proofs of the propositions in the text are contained in our 1977 working paper.

plausible to postulate that the subjective expected benefits of political participation to the ith gainer, net of his contributions, increase in the total contributions by all gainers to the lobby; decrease in the total contributions by losers to a free trade lobby; increase in the stake which i has in the outcome; and are related to i's contribution (above and beyond its effect on the total level of contributions) in the following way: the net gain is assumed to increase and then decrease in i's contribution. This captures the peer pressure of not contributing at all (the "noticeability" factor in Olson, p. 49n), nonpecuniary personal satisfaction, and Veblen-type vanity effects. The marginal gain functions of individual i are continuous, decreasing in own and total contributions and increasing in i's stake. We then make some very general assumptions about how the political process balances gainer contributions, loser contributions, and social costs in setting expected tariffs, and assume that the net gain functions of the individuals in each lobby are additively separable in the total contributions by the opposing lobby.

This framework guarantees the existence and uniqueness of a noncooperative equilibrium among the n contributors to the gainers lobby and the m contributors to the losers lobby, and has the following properties: 1) Many of the players may not contribute, even though they have a stake in the outcome; if no one contributes to a given side, no formal lobby exists and the group is represented only through its votes. 2) There is an endogenous critical minimum value of the stake of individual i, below which he contributes nothing as his stake varies and above which his contribution increases with his stake. 3) For the other $n-1$ gainers who contribute, their contributions are unaffected by increases in i's stake if he is not a contributor and decrease if he is a contributor. In the latter case, i's increased contribution exceeds the decrease of the others so that total contributions to the lobby increase with the increase in i's stake. This is the Olson effect of the small exploiting the large.

A fourth result deals with another form of the small exploiting the large, namely, the Olson-George Stigler proposition that an increase in the concentration of benefits, holding the total stake of the gainers constant, will increase total contributions. Assume that the stake of gainer 1 exceeds that of gainer 2. Let us increase the concentration of benefits by increasing 1's stake by $1 and decreasing 2's stake by $1. If total contributions to the lobby increase, it must be true that gainer 1's marginal propensity to contribute with respect to the stake is higher than gainer 2's. Within a given industry, these authors assume identical functions across agents relating contributions to stakes. Hence, their proposition incorporates implicitly the strong assumption that individuals have an increasing marginal propensity to contribute to lobbying with respect to the stakes. When Stigler turned to his data he was perplexed that he found "some hint of a negative [rather than a positive] correlation of association resources and concentration ratios" (p. 364–65). While a negative relationship refutes the traditional way of viewing the problem, it may not be implausible since, in our framework, it is consistent with a decreasing rather than an increasing marginal propensity to contribute.

Stronger results can be derived from an explicitly Olson-type formulation in which the total receipts by the lobby increase linearly in the stake of the representative member of the gaining group and nonlinearly in his share of total contributions. The coefficients on these terms capture what has come to be called the perceived effectiveness and noticeability of the participants, respectively. This model provides a framework for analyzing the question of whether consolidating political units increases or decreases the power of narrowly based but strong (protariff) interest groups relative to broader but less effective (consumer) groups. For example, is a tariff lobby more or less powerful as we move from races for the U.S. House of Representatives to the Senate to the

presidency? The results can go either way. The answer hinges on the size of the perceived effectiveness coefficient for protectionists relative to free traders. For example, the total stakes (for both sides) may be twice as high in a Senate race relative to a House race because the senator represents twice as many import-competing firms and consumers as the member of the House. But protectionist contributions may increase relative to free trade contributions in moving from the House to the Senate race if the protectionists' lobbying coefficient on the stakes is larger than the same coefficient for the free traders (because of greater success in overcoming the free-rider problem).

Finally, consider the question of how a lobby or a rent seeking individual should contribute in a two-man political campaign. Should it given to neither, one, or both politicians? To answer the question, we must make an assumption about how the lobbies and the politicians interact. There are three possibilities: neither considers the behavior of the other in making its decision; the lobbies contribute subject to an expected reaction function by all politicians; or each politician quotes his policy position (say, on the tariff) to maximize his probability of election, internalizing the reaction functions (i.e., flows of money) of the lobbies. For several reasons, we choose the latter as the most useful. This means that as each lobby contributes, the equilibrium tariff quoted by the politician does not change since he has already anticipated the flow of funds. This abstracts from any access effect.

In this situation, the only variables under the lobby's control are the probabilities of election of the politicians, say, politicians 1 and 2. Assume that man 1 quotes a 20 percent tariff as his best strategy while man 2 chooses 5 percent (for simplicity, the interpolitician equilibrium is assumed to Cournot-Nash). If the lobby is not large enough to have any effect on the outcome of the election, it should contribute to neither side. If it is large enough to have an effect, it should contribute to, at most, one side (the "campaign contribution specialization theorem"): for example, protectionists should contribute to the high-tariff politician. This holds regardless of whether the high-tariff politician has a high or low probability of election. The protectionists should invest in politician 1 up to the point where the marginal effect of the last dollar contributed on man 1's chances of election times the dollar value of the *difference* in the value of the two politicians to the protectionists just equals the last dollar given (i.e., the marginal political revenue equals marginal cost). If the two politicians quote the same tariff, there will be no difference in their economic value to the protectionist lobby and it should give to neither. The logic of large contributors giving to at most one side is clear. Since the contributor can only manipulate the probabilities of election and these sum to 1, contributing to both sides is self-defeating. Data from a sample of individual contributions to the 1972 presidential race between Nixon and McGovern is not inconsistent with this result: contributors who give to only one candidate were significantly larger than those who gave to both. However, the results of this paragraph should be viewed with skepticism because the data are crude and the access effect is ignored.

II. The Politicians

Each politician who maximizes his probability of election must calculate how campaign contributions and voters will respond to alternative tariff levels before he announces his position. In many cases, including some tariffs, there is a well-organized but narrowly based coalition for the special interest legislation and a large but disorganized coalition against it (for example, consumers). Since the latter frequently generate no campaign contributions, the only constraint on the level to which the high-tariff politician (hereafter, politician 1) will raise his proposed tariff in a two-man race is a negative general voter effect. This incorporates accusations of having been "bought off" by his opponent

and the negative effects of the tariff on consumer's surplus (both the redistribution and the deadweight loss effects).

In this situation, politician 1 will perform the following experiment: he will raise his tariff until the positive effect of increased funds on his probability of election is just offset by the negative general voter effect. This is the tariff he announces. What is the likely reaction of politician 2? Should he quote a zero tariff to maximize the negative voter effect against 1? Probably not. If there are no small donors and no free trade lobby, we know from the campaign contribution specialization theorem that all of the funds will go to 1. But *how much* goes to 1 depends on the difference in the tariff levels between 1 and 2. Politician 2 may not quote a zero tariff: if he lowers his quoted position too far, the negative effect on 2 of 1 getting more funds more than offsets the gain to 2 from the free trade voter effects.

Consider next the general case in which lobbies are not ruled out for either side. The probabilities of election for politicians 1 and 2 are determined by their contributions and their tariff positions. But contributions are a direct function of the tariff quotations so that the probabilities of election can be written solely as a function of the tariff positions, T_1 and T_2. In a simple interior Cournot-Nash equilibrium between the two politicians in T_1 and T_2 space, we find that the slopes of the reaction functions have opposite signs (the "reverse slope theorem"). This emanates from the second-order conditions of politician 1 choosing T_1 to maximize his chances and politician 2 choosing T_2 to minimize 1's chances. The implication is that near equilibrium, politician 1, say, emulates the actions of 2 (when T_2 rises, T_1 is increased), while 2 counteracts the actions of 1 (when T_1 rises, T_2 is reduced).

The classic Hotelling solution of the politicians quoting identical tariff positions is a special case in this analysis. It is guaranteed only when the two politicians are identical in every respect: otherwise the reaction functions cannot be expected to intersect along the 45 degree line in T_1, T_2 space. If the politicians were different but the political equilibrium occurred along the 45 degree line, we know from the campaign contribution specialization theorem that *large* contributors would sit out the race (we do not attempt to model small contributors).

Consider an increase in the power of the tariff lobby brought about by, say, a decrease in the lobby's organization costs. Assume that the reaction functions of both politicians shift out from the origin, indicating a preference by both, *ceteris paribus*, for higher tariffs. Since one of the functions has a negative slope, there are cases in which one of the politicians will increase his tariff while the other will lower his (the "reverse shift theorem").

The reverse shift theorem suggests another result. The expected tariff before the election, $\bar{T}$, is an average of T_1 and T_2, with the probabilities of election of politicians 1 and 2 as weights. There are plausible situations in which the increased power of the tariff lobby will lead to a decrease in $\bar{T}$. But the lobby can still be better off since expected net political revenue is up: costs fall more than revenue. A parallel is a decrease in costs in an industry with elastic demand: gross revenue falls as price falls but profits rise because of decreased costs.

While these results are untested, they have implications for empirical studies of tariff setting. Assume that we could rank all 435 *U.S.* congressional seats from least protectionist to most protectionist. The reverse shift theorem suggests that as we move to more protectionist districts, T_1, say, can rise and T_2 may fall. If this is the case, the variance in the tariff quotations would increase, providing that moving to more protectionist districts is accompanied by the high-tariff politician raising his tariff (this case seems more plausible than the reverse, although stability does not require it). This situation also raises the possibility that movement to more protectionist districts may be accompanied by a decline in the mean tariff, $\bar{T}$.

REFERENCES

W. A. Brock and S. P. Magee, "An Economic Theory of Politics: the Case of Tariffs," mimeo., May 1974.

——— **and** ———, "The Economics of Pork-Barrel Politics," Center Math. Stud. Bus. Econ., rept. 7511, Univ. Chicago, Feb. 1975.

——— **and** ———, "The Economics of Special Interest Politics: The Case of the Tariff," work. paper 78-01, Bur. Bus. Res., Univ. Texas-Austin 1977.

——— **and** ———, "International Trade and Tariffs: the Economics of Special Interest Politics," mimeo., 1978.

James E. Cairnes, *Some Leading Principles of Political Economy*, London 1874.

S. P. Magee, "Three Simple Tests of the Stolper-Samuelson Theorem," work. paper no. 77-28, Bureau Bus. Res., Univ. Texas-Austin, Feb. 1977.

——— **and William A. Brock**, "The Campaign Contribution Specialization Theorem," mimeo., Feb. 1976.

Mancur Olson, *The Logic of Collective Action*, Cambridge 1965.

G. J. Stigler, "Free Riders and Collective Action: An Appendix to Theories of Economic Regulation," *Bell J. Econ.*, Autumn 1974, *5*, 359-65.

W. Stolper and P. A. Samuelson, "Protection and Real Wages," *Rev. Econ. Stud.*, Nov. 1941, *9*, 58-73.

Understanding Collective Action: Matching Behavior

By JOEL M. GUTTMAN*

This paper develops an approach to understanding voluntary collective action. A simple model illustrating this approach predicts Pareto optimal provision of a nonexcludable public good in the case of identical actors with perfect information, regardless of the number of actors. In this approach, actors voluntarily subsidize each other's contributions to the provision of a public good. Each actor individually finds it optimal to match other actors' contributions dollar for dollar, and this matching behavior leads to a Pareto optimal outcome from the viewpoint of the group as a whole.

The approach developed here differs from two other sets of proposed solutions to the "free-rider" problem. One set of proposed solutions, which may be called "coercion solutions," simply assert that individuals are forced to contribute toward the provision of collective goods, once desired quantities of such goods are known (for example, Mancur Olson; Gary Becker's "theory of collusion;" Theodore Groves and John Ledyard). These solutions, however, beg the question of how the coercion itself is financed, since the policing of collective agreements is itself a public good: noncontributors cannot be excluded from benefiting from the public good resulting from the coercion.

Other proposed solutions of the problem of voluntary collective action assume some special property of the public good. Olson's "by-product" solution assumes that the public good can be jointly produced with a private good, and that the private good cannot be produced as cheaply without also producing the public good. George Stigler's "asymmetry" solution, as a second example, assumes that individuals have differing interests regarding the exact form that the public good will take, leading them to contribute so that the good that is provided is optimal from their own individual viewpoints. Both of these solutions implicitly introduce some form of "private-ness" into the public good whose provision they try to explain.

This paper avoids limiting assumptions of special characteristics of public goods, and also does not postulate any coercion in the provision of the public good. After a description of the model, some examples of the predicted matching behavior and experimental evidence are briefly discussed.

I. The Model

A. *The Process of Collective Action*

The process of collective action is viewed as consisting of two games: 1) a game in which "flat contributions" by actors to the provision of public goods are determined (the a game), and 2) a game in which matching rates are determined (the b game). Each actor's final contribution toward the provision of the public good is his flat contribution, plus his matching rate times the sum of the other actors' flat contributions:

$$(1) \qquad x_i = a_i + b_i \sum_{j \neq i} a_j$$

x_i = actor i's contribution

a_i = his flat contribution

b_i = his matching rate

The payoff of actor i, π_i, is a (concave) function of the combined contributions of

*The University of California-Los Angeles and The Rand Corporation. I wish to thank H. Demsetz, P. Dubey, J. W. Friedman, J. Hirshleifer, M. Stutzer, E. Thompson, and particularly J. Riley, L. Shapley, and L. Kotlikoff for helpful comments on earlier drafts of this paper. H. Salehi and I. Maoz assisted in conducting the experiments described in this paper.

all actors, x, minus the cost of i's own contribution:

$$(2) \qquad \pi_i = f_i(x) - x_i$$

Thus π_i is a function of a_i, b_i, and the other actors' a_j and b_j, which jointly determine the amount of the combined contributions x, as well as i's own contribution x_i.[1]

In order to determine his flat contribution, a_i, actor i must first know the effective price of the public good, which for him is determined by the sum of the other actors' matching rates. This can be seen most clearly by writing out π_1 when there are two actors:

$$\pi_1 = f_1[a_1(1 + b_2) + a_2(1 + b_1)] - a_1 - b_1 a_2$$

The first-order condition of optimal a_1 is

$$(3) \quad \frac{\partial \pi_1}{\partial a_1} = f_1'(x)(1 + b_2) - 1 = (1 + b_2)\left[f_1'(x) - \frac{1}{1 + b_2}\right] \leq 0$$

with $a_1 = 0$, if the inequality is strict. Thus actor 1 increases his flat contribution, a_1, until the marginal benefit of the public good $f_1'(x)$ equals the effective price he faces, $1/(1 + b_2)$. If x is so large that this marginal benefit is less than the effective price, he sets his flat contribution equal to zero. Therefore, actor 1 must know the other actor's matching rate b_2 before setting his own flat contribution.

The matching rates thus are "data" in determining the flat contributions. But the latter are not data in determining the matching rates, because the very purpose in matching other actors' contributions is to increase those contributions. Therefore, the game determining the matching rates (the b game) precedes the game determining the flat contributions.

In the b game, actors are assumed to have the following information: knowledge of the other actors' gross benefit functions: and knowledge of their matching rates, which are taken as given. With this information, each actor can predict the outcome or possible set of outcomes of the upcoming a game for whatever matching rate he chooses. The actor maximizes his payoff, taking into account the effect of his choice of matching rates on the outcome of the a game.

The a game, like the b game, is viewed as a Nash noncooperative game. In the a game, each actor takes the other actors' flat contributions (as well as the previously determined matching rates) as given and maximizes his payoff by varying his own flat contribution a_i. The a_i will be chosen so that the total quantity of x is individually optimal for each actor choosing a positive a_i. Once the equilibrium a_i are determined, the payoffs of all actors are determined by equations (1) and (2).

Both games are viewed as one-period rather than multiperiod games, and, as indicated above, both are Nash noncooperative games. While the Nash assumption assumes naivete on the part of the actors, a more realistic level of sophistication by the actors is built into the model, by allowing actors to match other actors' contributions and to take others' matching rates into account in determining their own contributions.

B. *An Example*

Suppose there are two identical actors, John and Karl, each with the payoff function

$$(4) \quad \pi_i = x^{1/2} - x_i, \quad i = j, k$$

Substituting the definitions of x_j and x_k into (4), we have

$$\pi_j = [a_j(1 + b_k) + a_k(1 + b_j)]^{1/2} - a_j - b_j a_k$$

[1] The payoff function (2) ignores effects of changes in one actor's contribution on other actors' wealth (see John Chamberlin; Martin McGuire). In effect, the actors are viewed as profit-maximizing firms, purchasing the public good as an input to their production from a "vendor" at a constant price of unity. The production of the public good is exogenous to the model. Moreover, the public good is assumed to be perfectly divisible in production.

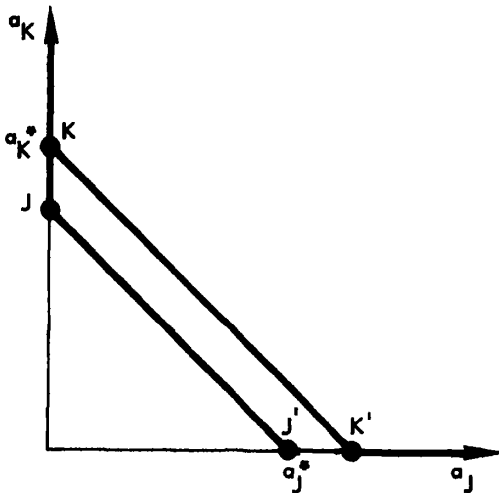



FIGURE 1. EQUILIBRIUM IN THE a GAME, $b_j > b_k$

John will set his flat contribution so that

(5) $$(1/2)x^{-1/2} \leq 1/(1 + b_k)$$

Similarly, Karl will satisfy

(5′) $$(1/2)x^{-1/2} \leq 1/(1 + b_j)$$

Strict equality can obtain in only one of these two expressions if $b_j \neq b_k$. If, for example, $b_k < b_j$, only (5′) will be an equality—i.e., Karl will find an interior solution of a positive flat contribution- while John will set his flat contribution equal to zero. Geometrically, if we draw *reaction functions* for the a game as in Figure 1, the slopes of the reaction lines will be equal, and the reaction line of the player facing the higher matching rate (Karl) will lie outside that of the player facing the lower matching rate (John). John's equilibrium flat contribution will then be zero.

Thus one possibility is that if Karl sets his matching rate in the b game to exceed John's matching rate, Karl can then expect that his own flat contribution in the a game will equal zero. Intuitively, John will have the greater demand for the public good, and will pay the full flat contribution necessary to satisfy this demand. Thus Karl's payoff, in this case, would be

(6) $$\pi_k = x_j^{*1/2} - b_k a_j$$

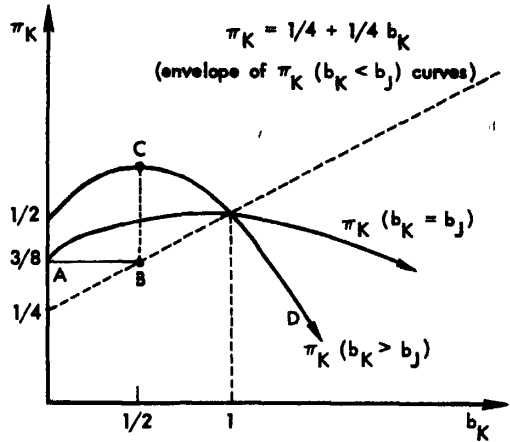



FIGURE 2. PAYOFF CURVES WHEN $f(x) = x^{1/2}$

That is, his only cost would be matching John's flat contribution. Here x_j^* denotes John's individually optimal x, given b_k, which is the quantity of x satisfying (5) when (5) is a strict equality. By solving (5) for x_j^*, we obtain $x_j^* = (1/4)(1 + b_k)^2$. Moreover since in this case, $a_k = 0$, we have $x = a_j(1 + b_k)$ or $a_j = x_j^*/(1 + b_k)$. Substituting these expressions for a_j and x_j^* into (6), we have

(6′) $$\pi_k \Big|_{b_k > b_j} = 1/2 + (1/4)b_k - (1/4)b_k^2$$

A second possibility is that Karl sets his matching rate below John's. By an analogous derivation, one can show that in this case,

(7) $$\pi_k \Big|_{b_k < b_j} = (1/2)(1 + b_j) - x_k^*/(1 + b_j) = 1/4 + (1/4)b_j$$

In Figure 2, $\pi_k(b_k < b_j)$ is shown for $b_j = 1/2$ (Segment AB); the dashed line is the right-hand envelope of similar horizontal lines parameterized for different b_j's.

The third possibility is that Karl set his matching rate equal to John's. In this case, the two reaction lines in Figure 1 would coincide, and the equilibrium flat contributions would be indeterminate. It turns out, however, that whatever the equilibrium flat contributions are, the payoff curve plotting b_k into π_k, given $b_j = b_k$, passes through

the intersection of the payoff curve drawn under the assumption $b_k > b_j$ and the envelope of the payoff curves drawn under the assumption $b_k < b_j$. (See Figure 2.) Moreover, it can be shown that for any $b \neq 1$, the $\pi_k(b_k = b_j)$ curve lies between the $\pi_k(b_k > b_j)$ curve and the envelope of the $\pi_k(b_k < b_j)$ curves.

The joint intersection of the three payoff curves determines the equilibrium matching rates of each player, which are unity. For example, suppose that, initially, John sets his matching rate equal to 1/2, attempting to reach the maximum of the $b_j > b_k$ curve. Then the payoff curve for Karl is *ABCD* in Figure 2. Karl will set b_k to exceed b_j, since the strategy $b_k > b_j$ dominates the other two strategies in the region of $b_k = 1/2$. John, in turn, will set b_j to exceed b_k, and so on, until both players find that the strategy of setting the higher matching rate no longer dominates the other two strategies. Suppose, alternatively, that initially John's matching rate exceeds unity, so that Karl's optimal strategy is to set $b_k < b_j$. John, having identical π curves, would do the same, until once again both matching rates would equal unity.

When both matching rates equal unity, a Pareto optimal provision of the public good is assured, because then the effective price of the good facing each actor is 1/2. Each actor would then set $f'(x) = 1/2$. Thus the *sum*, $2f'(x)$, would be unity, the "nominal" price of the good—that is, the sum of the marginal benefits would equal the price of the public good, which is the Samuelsonian optimality condition.

The intuitive basis for the equilibrium matching rate of unity is that at any other rate, each actor can profit from increasing or decreasing his matching rate above or below that of the other actor. If, for example, John increases his matching rate above Karl's matching rate, he can reduce his flat contribution to zero. Karl would then increase *his* flat contribution so that the sum of the flat contributions remained constant, in order that x remain at its individually optimal level. John would then have to match this increase in Karl's flat contribution—which could leave him no better off than before, if his matching rate were unity. If his matching rate were initially less than unity, however, he would gain by setting the higher matching rate and reducing his flat contribution to zero. Conversely, if the matching rate were greater than unity, he would gain by setting the lower matching rate. Only if both matching rates were unity, would there be no incentive to change one's own matching rate.

C. *Implications of the Model*[2]

With n identical actors, the prediction of Pareto optimality remains correct regardless of the $f(x)$ functions, assuming these are strictly concave. This can be verified by comparing $\pi_k(b_k > b_j)$, $\pi_k(b_k = b_j)$, and $\pi_k(b_k < b_j)$, with b_k of the $b_k > b_j$ strategy being equal to b_j of the $b_k < b_j$ strategy. It can be shown that, in general,

$$\pi_k\Big|_{b_k > b_j} \gtreqless \pi_k\Big|_{b_k = b_j} \gtreqless \pi_k\Big|_{b_k < b_j}$$

as $b_k \lesseqgtr 1$. With identical actors, then, Figure 2 generalizes to depict the determination of the equilibrium b_i regardless of the form of $f(x)$, as long as $f(x)$ is strictly concave. This equilibrium, moreover, will be unique. With many actors, b_j denotes the common matching rate to which all actors j will tend to converge.

With two nonidentical actors, there continues to be a unique Pareto optimal equilibrium. The actor with the larger marginal benefit function chooses a higher matching rate than the other actor. With more than two nonidentical actors, some indeterminacy emerges in the equilibrium, and inefficient equilibria become possible. The divergence from Pareto optimality predicted by the model, however, is considerably smaller than that which would be predicted if actors simply took other actors' total contributions as given.

II. Some Empirical Evidence

There are real world examples of explicit

[2]Proofs of statements contained in this section can be obtained from the author on request.

matching behavior such as that described by the model. The federal government matches public investments by state governments to finance such public goods as highways, individuals make contributions to charities contingent on matching by others, and so forth. More generally, whenever individuals voluntarily agree to split the cost of some collective good, they are implicitly matching each other's contributions. The fact that no outside enforcement is necessary, in many instances, to enforce such agreements supports the model.

More systematic evidence supporting the model can be found in certain "Prisoner's Dilemma" experiments. Lester Lave confronted subjects with robots, one of which played a "Khruschev pattern" in which noncooperative behavior was interspersed with random lapses of cooperative behavior. The subjects' response was to match the cooperative lapses with cooperative behavior, until they learned that the cooperative lapses were truly random, at which point they stopped matching. Austin Hoggatt performed a similar experiment with robots, and also found that subjects match cooperative behavior of robots.

I have performed experiments *without* robots, in which physically isolated subjects had the option of keeping an endowment of $7 given to them at the beginning of the experiment, or contributing part or all of it to a "common pool." The players had information of the contributions of other subjects in the previous "rounds" of the game. The payoffs to subjects were identical concave functions of the size of the common pool, minus the amount of the subject's contribution. Matching rates were estimated by regressing current contributions on the combined contributions of other actors in the previous round. These were estimated as being positive, on the average, when the number of subjects was three, but not when the number of subjects was six. Combined contributions, however, were larger with the larger group size and were larger than what would be predicted if each actor acted in isolation. Estimated matching rates with six players suggest the importance of a factor neglected by the model—computation and observation of matching rates is costly, and these costs may increase with group size, diminishing the returns to matching behavior.

III. Concluding Remarks

The model's prediction of Pareto optimal provision of public goods with identical actors and perfect information is reminiscent of, but considerably stronger than, the Coase Theorem. The latter asserts that actors can attain socially optimal outcomes through explicit negotiations and enforcement of property rights. The process described in this paper involves no negotiations or enforcement of property rights, but nevertheless yields a Pareto optimal provision of public goods, given perfect information and identical actors.

REFERENCES

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *76*, 169–215.

J. Chamberlin, "Provision of Collective Goods as a Function of Group Size," *Amer. Polit. Sci. Rev.*, June 1974, *65*, 707–16.

T. Groves and J. Ledyard, "Optimal Allocation of Public Goods: A Solution to the Free-Rider Problem," *Econometrica*, May 1977, *45*, 783–809.

A. C. Hoggatt, "Measuring Behavior in Quantity Variation Duopoly Games," *Behav. Sci.*, Mar. 1967, *12*, 109–21.

L. B. Lave, "Factors Affecting Cooperation in the Prisoner's Dilemma," *Behav. Sci.*, Jan. 1965, *10*, 26–35.

M. McGuire, "Group Size, Group Homogeneity, and the Aggregate Provision of a Pure Public Good Under Cournot Behavior," *Publ. Choice*, Summer 1974, *18*, 107–26.

Mancur Olson, Jr., *The Logic of Collective Action*, Cambridge, Mass. 1965.

G. J. Stigler, "Free Riders and Collective Action: An Appendix to Theories of Economic Regulation," *Bell J. Econ.*, Autumn 1974, *5*, 359–65.

Voters, Legislators and Bureaucracy: Institutional Design in the Public Sector

By MORRIS P. FIORINA and ROGER G. NOLL*

The purpose of this paper is to outline a theory of representative democracy which explains why rational actors construct an excessively bureaucratized government. We define excessive bureaucratization as the selection of an inefficient production technology for the public sector, characterized by relative factor proportions that entail more bureaucracy than the proportions that would minimize total costs. Thus, the question of excessive bureaucracy is related to but conceptually different from whether a particular policy is worthwhile. Furthermore, it presumes a concern more fundamental than the observation that implementing a public policy inevitably requires the expenditure of scarce resources.

Section I describes a theory formally presented in the authors' forthcoming article. Section II develops the predictions of the theory, most of which have not been tested. Section III outlines reforms that might undo some of the effects that the theory predicts.

I. The Theory: A Voter's Dilemma

We assume that in choosing among alternative political actions, voters, bureaucrats, and politicians pursue their self-interests. For voters, this means casting votes in a manner that maximizes expected utility, given the platforms of competing candidates. For bureaucrats, this means maximizing some measure of the size of the bureaucracy. For politicians, this means maximizing the probability of election.

Politicians can affect the welfare of individual voters in three ways. First, some of the arguments of utility functions are government activities. Second, government redistributes income through taxation, subsidization, and expenditures on the production of public goods. Third, government bureaucracies, in carrying out public policies, impose costs on citizens by ensnarling them in red tape.

We assume that government activities can be characterized by a production function, the arguments of which can be usefully classified into bureaucratic and nonbureaucratic inputs. Bureaucratic activities include keeping formal records, developing and enforcing procedures to govern relations between the bureau and its clients, communicating among parts of the organization, and controlling and evaluating personnel. These activities impose an external cost on citizens because their complexity creates an informational problem for citizens who seek services, because the data required by the bureaucracy come in part from the agency's clients who incur some expense providing them, and because bureaucratic processes are time consuming.

A legislator serves the home constituency in various ways (see Fiorina, pp. 41–49). Legislators collectively decide general issues of public policy by majority rule votes. (We assume that the distribution of voter preferences supports a majority rule equilibrium. While unnecessary, this assumption simplifies our argument.) Each legislator also is a near-monopolistic supplier of unpriced facilitation services to constituents. Facilitation services take several forms: intervening in bureaucratic processes to aid citizens ensnarled in red tape, providing information to citizens who want to know how and where to approach the bureaucracy, and acquiring for constituents a share of "distributive" activities. A government activity is distributive if it is divisible into subactivities, each of which is evaluated and decided upon separately and is beneficial to a rela-

*California Institute of Technology.

tively small proportion of the electorate. Examples are federal construction projects, categorical grant programs, and commodity-specific tariffs.

A bureaucracy can assist a legislator in carrying out facilitation. It can accommodate inquiries by the legislator on behalf of constituents by providing information about certain services or by expediting a decision. It can propose, and try to justify, distributive activities in a legislator's home district. A rational bureaucrat will use these possibilities to serve the objectives of the bureau by rewarding legislators who support its programs and appropriations.

If a bureaucracy responds favorably to facilitation activities, it makes legislators more attractive to their home constituencies. Effective facilitation lowers the external costs of bureaucracy and raises the share of government distributive activities to the constituency. The latter is attractive because the taxes used to finance a project in one district are imposed on everyone, whereas the benefits are concentrated. Moreover, performance as a facilitator depends on the personal actions of the legislator, enabling a legislator to claim credit for it (see David Mayhew, pp. 52–59). In contrast the public policy decisions of the legislature are unlikely to be affected by the vote of a single legislator.

One consequence of the preceding argument is that legislators and bureaucrats have an incentive to provide government services in an excessively bureaucratized manner. To do so raises the demand for facilitation services. The electoral process does not check this tendency because voters face a prisoner's dilemma in choosing among candidates. If voters disapprove of excessive bureaucratization, electing a legislator who attacks bureaucratic inefficiency will be unlikely to alter the outcome of a majority-rule legislature, but will produce a less effective facilitator.

As the public bureaucracy grows larger, the importance of the performance of facilitation will grow, and a legislator who is a good facilitator will be increasingly likely to be reelected. A challenger who is unproven as a facilitator is a riskier choice than an effective incumbent, and consequently provides a lower expected payoff in this role. This tendency will be accentuated if a legislator becomes a more effective facilitator over time. Because part of facilitation is, the possession and use of information which is acquired through experience, and because seniority enhances the influence of a legislator in determining the fate of an agency, incumbents can be more effective facilitators than their challengers.

II. Applications: Evidence and Predictions

Direct observation of the production function for a government activity is especially difficult, and a test of the primary implication of the theory—that public activities are excessively bureaucratized—is beyond us at present. We offer some indirect evidence, and present other predictions of the theory.

A. *Facilitation and Congressional Elections*

During the early postwar period congressional elections were low information affairs in which most citizens voted according to traditional partisan affiliations that reflected a generalized preference for one party (see Donald Stokes and Warren Miller). Since the early 1960's, the impact of partisan affiliations on congressional votes has been declining. The most important influence to take up the slack is incumbency. The "incumbency advantage" now appears to be 5 to 10 percent in House elections (see Robert Erikson; John Ferejohn) and a few points higher in the Senate (see Warren Kostroski). The incumbency advantage is not a result of post-Wesberry redistrictings favorable to incumbents, nor to increased knowledge of incumbents—despite their greatly increased advertising before a more educated, less partisan electorate (see Fiorina, ch. 3).

Elsewhere (see the authors, 1977) we show that rational incumbents should base reelection efforts on facilitative activities rather than programmatic advocacy. This conclusion follows from the relatively un-

controversial nature of the former and the greater personal effectiveness of the legislator in facilitative activities.

Effective facilitation requires resources, and in the past two decades these resources have increased dramatically. Congressional employment is one example. Several support organizations have been established (Legislative Reference Service, Office of Technology Assessment, Congressional Budget Office). Committee staffs have expanded, and the personal staffs of congressmen have grown from six to a baker's dozen (see Table 1). The ostensible reason for these developments is the growing complexity of governance. Congressmen also argue that this growth offsets an imbalance of expertise between the executive and legislative branches. Perhaps, but another reason may be growth in facilitation activities.

Few solid data exist, but congressmen have greatly expanded their district staff presence (see Table 1), and presumably the staff in Peoria has less to do with legislation than with facilitation. Moreover, rough estimates indicate that even the Washington staff spends more than half its time on facilitation (see Fiorina, p. 59). Other innovations that serve the facilitation role are mobile district offices, allowances for computerized records about constituents, and more paid trips home (representatives had three in 1960 and thirty-two in 1978).

TABLE 1—THE PERSONAL STAFFS OF CONGRESSMEN

	1960	1967	1974
Total staff	2,344	3,276	5,109
Percent assigned to district	14	26	34
Percent district offices open only when congressman is home or after adjournment	29	11	2
Percent congressmen with multiple district offices	4	18	47

Source: Annual Congressional Staff Directories, compiled by Charles B. Brownson.

B. *Policy Trends*

Three general implications about the construction and implementation of public policy follow from the theoretical discussion in the first section.

First, because legislators profit from effective facilitation and bureaucrats profit from accommodating legislators, both have an incentive to structure programs so that facilitation is important. A principal way to do this is to inject distributive elements into a program. Expenditure programs, for example, can employ project-by-project decisions or automatic distribution according to a fixed formula. One explanation for Lyndon Johnson's wizardry in steering Great Society programs through Congress was his willingness to use distributive politics to purchase congressional support. In 1964, over one-third of all federal grant programs allocated funds by formula. During 1965 and 1966, the number of grant programs increased by 70 percent, and only about one-sixth of the new programs used allocation formulas (see Advisory Commission on Intergovernmental Relations, p. 151).

Congress can encourage facilitation in ways other than distributive policy making. For example, regulatory legislation can be so vague that an agency must make numerous detailed decisions before implementing it (see Theodore Lowi). Later, oversight subcommittees find it easier to use budgetary review as a lever for affecting regulations than to change a law.

A second implication concerns the increasing acceptance of new programs after they are enacted. While legislative proposals may be controversial, opposition will decline once the program is established. Before enactment no facilitation takes place. If constituents care about a proposed program, they evaluate politicians according to their positions on the issue. Once the program is established, both supporters and opponents need facilitation. A legislator's policy position becomes less important than facilitative abilities. Die-hard legislators who oppose an agency unsuccessfully only succeed in penalizing their districts, perhaps by overt agency actions, but more likely by foregoing what would be their "due" under the

program (see Barry Weingast). Republicans who vote en bloc against new social programs quietly go along with reauthorization (see David Stockman). We suspect that new programs gradually become altered in the direction of distributive politics, which would provide a further reason why programs that initially trigger major political battles gradually become the object of a political consensus.

A third policy implication is the loosening of policy ties between representatives and their constituencies. Because facilitation gives incumbents an electoral advantage, their policy positions are less important in their constituents' evaluation of them. For any given incumbent advantage arising from facilitation, an incumbent may deviate from the position of the median voter by some amount and still expect to defeat a challenger who adopts the median voter's ideal position (see the authors, forthcoming).

The model does not imply that the amount of public goods provided by the government is greater than would be the case in the absence of facilitating activities. Because the public sector adopts inefficient production technologies, the amount of public goods that the median voter desires will be less than if production were at minimum cost; however, the incumbency advantage may offset this effect if the incumbent's policy position moves in the direction of more public goods than the median voter prefers. The incumbent has a generalized incentive to move in this direction because a larger public sector implies a greater demand for facilitation and a correspondingly greater incumbency advantage. Additionally, to move outside the model for a moment, as the policy ties between incumbents and their districts weaken, the former are increasingly at liberty to cooperate with special interests that desire some particular government activity or expenditure.

To summarize, we attribute the increase in the incumbency advantage in congressional elections to the gradual transformation of congressmen from makers of national policy to ombudsmen and grantsmen. The scope of the federal government has expanded during the past two decades, creating greater opportunities for citizens to profit from bureaucratically administered programs, and numerous occasions for citizens to run afoul of bureaucratically promulgated regulations. An incumbent's experience and seniority are an important resource which disappears upon election of a challenger. In this brave new world, citizens have come to attach more importance to the facilitation activities of congressmen (see Roger Davidson).

III. Prospects For Change

The foregoing theory of the legislative process does not work in quite the way that constitutional theory postulates. Our theory simply isolates the incentives facing modern voters, bureaucrats and legislators, incentives created by the institutions within which legislators and bureaucrats act. To change the system requires changing the institutions. Several possibilities are imaginable; none appear likely. From most to least drastic they include:

1. *Remove incumbents' facilitative powers*. This involves slashing staffs and removing the constitutional basis for congressional power over the bureaucracy, making the latter more responsible to the president, who has a national constituency. Of course, this possibility is fraught with the dangers of the imperial presidency.

2. *Change the electoral system for congressmen*. If legislators increasingly are elected for facilitative efforts, legislators are less responsive for their policy positions and *no one is responsible* for legislative policy. Altering this situation probably requires abandoning the single member district. Proportional representation makes a candidate's election depend on the percentage of the vote received by parties nationally rather than each candidate's personal percentage in one district. This makes legislators more dependent on the policy position of the parties, but eliminates representation of particular constituencies.

3. *Lessen incumbents' facilitation responsibilities*. In other countries the

ombudsman role is performed by a special office, rather than by individual legislators. As rational actors American congressmen steadfastly resist suggestions to establish a federal ombudsman. Even mild proposals such as Henry Reuss's Administrative Counsel of the Congress receive quick execution. British MPs, whose facilitation powers and resources do not compare to the American congressman's, approved the creation of the Parliamentary Commissioner for Administration (ombudsman) only after they were made the communication link between constituents and the new office.

4. *Cumulative policy failure*. A long-run prospect for change is inherent in ever less efficient policies that impose ever heavier external costs on citizens. If the situation deteriorates sufficiently, the voter's prisoner's dilemma might be broken: the attempt to elect antibureaucratic legislators could become rational even if failure to elect a majority resulted in losses to the districts which elect them.

Despite the preceding emphasis on prospects for changing the legislative system, we hasten to emphasize that this system has positive aspects. Bureaucracy is permanent, and information about it is valuable. Some argue that facilitation is the only role modern legislators can perform well (see Samuel Huntington). To reach judgments of the costs and benefits of legislative facilitation, we must recognize that facilitation exists and include it in our analysis.

REFERENCES

R. Davidson, "Our Two Congresses: Where Have They Been? Where Are They Going?," paper presented at the Southern Political Science Association Meetings, New Orleans 1977.

R. Erikson, "The Advantage of Incumbency in Congressional Elections," *Polity*, Spring 1971, *3*, 395–405.

——, "Malapportionment, Gerrymandering, and Party Fortunes in Congressional Elections," *Amer. Polit. Sci. Rev.*, Dec. 1972, *66*, 1234–355.

J. Ferejohn, "On the Decline in Competition in Congressional Elections," *Amer. Polit. Sci. Rev.*, Mar. 1977, *71*, 166–76.

Morris Fiorina, *Congress—Keystone of the Washington Establishment*, New Haven 1977.

—— **and R. Noll**, "Voters, Bureaucrats and Legislators: A Rational Choice Perspective on the Growth of Bureaucracy," *J. Publ. Econ.*, forthcoming.

—— **and** ——, "A Theory of the Congressional Incumbency Advantage," soc. sci. work. paper no. 158, California Instit. Technology, Apr. 1977.

S. Huntington, "Congressional Responses to the Twentieth Century," in David Truman, ed., *The Congress and America's Future*, New Jersey 1965, 5–31.

W. Kostroski, "Party and Incumbency in Post-War Senate Elections: Trends, Patterns, and Models," *Amer. Polit. Sci. Rev.*, Dec. 1973, *67*, 1213–34.

Theodore Lowi, *The End of Liberalism*, New York 1969.

David Mayhew, *Congress: The Electoral Connection*, New Haven 1974.

D. Stockman, "The Social Pork Barrel," *Publ. Int.*, Spring 1975, *39*, 3–30.

D. Stokes and W. Miller, "Party Government and the Saliency of Congress," *Publ. Opinion Quart.*, Winter 1962, *26*, 531–46.

B. Weingast, "A Rational Choice Perspective on Congressional Norms," soc. sci. work. paper no. 142, California Instit. Technology, Oct. 1976.

Advisory Commission on Intergovernmental Relations, *Fiscal Balance in the American Federal System*, Vol. 1, Washington, Oct. 1967, 151.

DISCUSSION

DANIEL MCFADDEN, Yale University: While the papers in this session deal with various aspects of the problem of collective action, they have a common thread. They view the social decision as the outcome of a game with individuals, interest groups, legislators, or bureaucracies as players. In each paper, the goals and behavior of the players have been stylized in order to obtain clear-cut results. The authors are to be commended for obtaining meaningful analytic formulations of these problems, and simplifying the problems in ways that yield interesting first results. However, it is interesting to consider how these games might change if richer, and more realistic, strategies were available to the players.

Consider government bureaucracy as a player. The paper by Morris Fiorina and Roger Noll treats bureaucracy as a largely passive actor, evolving as a consequence of the election game played by congressional candidates. The only active role of the bureaucracy in this paper is to apply distributional penalties to opposition congressmen.

The serious and popular literature on bureaucracy suggests that it will be a more active participant, and that its strategies will be determined by the interplay of bureaucrats responding to the internal structure and incentives of the organization. Max Weber views the evolution of bureaucratic behavior as the result of individuals within the organization seeking power and avoiding personal liability, by replacing discretionary judgements with impersonal rules. Parkinson's laws suggests the mechanisms by which individual incentives are translated into the evolution of the bureaucracy. Kafka paints a bleak picture of the outcome of this process, and incidently points out the usefulness of a good facilitator.

Applying this view of bureaucracy to the world of Fiorina and Noll, one sees that while the form of bureaucracy is molded by legislation, the bureaucracy may itself become an active lobbyist for legislation affecting its future. It is particularly suited for this task because of its control of information and distributional power. Bureaucracies have a strong incentive to *not* yield distributional power to Congress, but rather to retain this power in order to develop their own constituencies and supporting lobbys, and to avoid exposure of individual bureaucrats to pressure which could create personal liability. In this view, excessive bureaucratization is in large measure a consequence of the internal incentives to individuals in the organization. The evolution of complex rules requiring facilitators occurs for Weberian and Parkinsonian reasons. As a player, the bureaucracy may actively attempt to maintain its own constituency, and to minimize the role of facilitators, and may be an active and effective lobbyist for its own growth.

If bureaucracy is an active player, then the games considered by Joel Guttman and by William Brock and Stephen Magee also change. In the Guttman world, the bureaucracy may actively seek an enforcement role for collective action, providing both a mechanism and a lobby for coercive solutions to problems of providing public goods. The preferences of government for quotas over tariffs, for environmental regulations over pollution prices, may be due more to the self-interest of bureaucrats than to the difficulty of understanding the theoretical advantages of decentralization. The Brock-Magee model also changes when bureaucracy becomes an active participant. In addition to buying a favorable policy position, campaign contributions now play a second role—the purchase of facilitation services, including favorable language in legislation establishing bureaucracies to enforce collective action. In this world, the lobbying strategies considered by Brock and Magee will be superimposed on "facilitation insurance" purchased from all viable candidates.

I have discussed government bureaucracy as one of the players in the collective action game whose role is more active, and more important, than the models treated in these papers suggest. I hope the evolution of this research area will

see the interesting initial results in these papers extended to more complex and realistic collective action games with richer behavioral models of each of the players.

Before closing, I have three specific comments on the papers. Fiorina and Noll suggest that direct test of the hypothesis of excessive bureaucratization are difficult. However, this may be possible where the public and private sectors provide the same or comparable services. Examples are public vs. private garbage collection, the post office vs. U.P.S., and Amtrak vs. the airlines.

The Brock-Magee paper assumes that voters face two candidates and one policy issue. In reality, there are many more issues than candidates, and each candidate represents a portfolio of policy positions. Voters then may have difficulty voting an issue, and candidates can play a game in which they can duck some issues and stress others. Lobbyists may then contribute to both candidates to get an issue suppressed, or to protect one candidate from attack by another. The contribution game is part of a supergame in which a politician may help a special interest group in many ways, from facilitation, logrolling, setting up straw men, etc.

Guttman's model assumes individualistic preferences—a reasonable starting point. However, individuals may care about not only the amount of the public good available, but also the level of effort. In a recent study of altruistic behavior, Steve Goldman has constructed examples where each consumer's willingness to contribute depends on the distribution as well as the contributions of others, and a Nash equilibrium fails to exist. In this case, the Guttman matching game may also fail to have an equilibrium.

WALLACE E. OATES, Princeton University: Last fall as a discussant in a session of the meetings of the American Political Science Association, I was struck by the fact that each of the papers could easily have been written by an economist. The interdisciplinary character of ongoing research in public choice is further in evidence in this session: papers on lobbying, bureaucratic behavior, and collective-choice mechanisms that might well have been the work of some political scientists. I find all this quite intriguing inasmuch as multidisciplinary efforts are often such a frustrating and disappointing enterprise. Yet here in public choice we have an instance of a very fruitful intersection of work. Both economists and political scientists, as a result largely of the momentum of their own efforts, have found themselves led to a similar approach and set of problems with a potentially quite rich yield.

The paper by William Brock and Stephen Magee is a most suggestive exploration of lobbying activities. Using a framework of profit-maximizing firms, utility-maximizing individuals, and vote-maximizing politicians, they generate a provocative series of propositions that are, in principle, at least testable. One can't get something for nothing in a complex issue like this, and so it is that these propositions come at the expense of some rather restrictive assumptions. In particular, I want to focus on their basic conception of a set of politicians who unequivocally announce their stands (here a tariff level) and presumably implement their positions if elected. It is this rather rigid view of political behavior that leads, for example, to their result that a lobbyist contributes to only one candidate: the one whose announced position is most favorable to the lobbyist's interests. I would suggest, however, as a first principle of lobbying behavior: "It is far better to have both candidates in your pocket than just one." And given a more flexible and realistic view which allows for back-stage deals and revised policies following upon election, I would conjecture that contributions to several candidates may well be the rational lobbying strategy.

The paper by Morris Fiorina and Roger Noll is of interest for its attempt to draw many political actors into a single conceptual framework: elected politicians, bureaucrats, and voters interact to determine a set of outcomes. However, I am uncomfortable with the explicit form of the

model and its implications. They distinguish between bureaucratic and nonbureaucratic inputs into public production. This I find a quite vague distinction, one which would be quite difficult to render operational. They use this framework to show tendencies toward excessive bureaucratization in the provision of public services; this takes the form of inefficient factor proportions or waste in the production process, largely to satisfy the politician's desire to provide "facilitation services."

All this is, however, quite unnecessary to explain waste in public production. The basic Niskanen model (in the case of demand constrained output) and some of its later, improved variants already imply such waste. There is inherent in the system of bureaucratic objectives and constraints a set of forces encouraging an excessive size of staffs and red tape. One need not invoke, as Fiorina and Noll do, a problematic kind of conspiracy theory to explain bureaucratic inefficiencies. Their contention is that politicians and bureaucrats conspire to create an impenetrable jungle so that the politician can get points (and votes) for leading members of his or her constituency through to the hidden pot of gold. There is no doubt some truth to this (they do perform a service by calling attention to facilitation services), but I think it a misplaced emphasis to base a theory of bureaucratic waste on this form of behavior. There are more important and pervasive forces inducing inefficiences in public agencies.

The paper by Joel Guttman is an interesting analysis that underlines usefully the forces at work promoting cooperation in the provision of public goods. However, I reject Guttman's contention that his is the first model to explain collective action in the absence of coercion. In fact, I think his analysis is basically in the spirit of the Lindahl model of voluntary exchange. Moreover, his results are fully consistent with the conventional wisdom on public goods: the likelihood of cooperation in the small-group case with such cooperation breaking down as the group becomes larger. His paper does, however, serve to emphasize the basic point that it is the presence of imperfect information and transactions costs that creates the public-goods problem and constitutes the *raison d'etre* for the public sector.

International Markets for *LDC*s— The Old and the New

By CARLOS F. DIAZ-ALEJANDRO*

> Within even the narrowest purview of the most abstract model of a competitive economy, efficiency requires public actions to deal with externalities, public goods, pervasive economies of scale, and incentives to destroy competition.

Who is responsible for such public actions in international transactions and markets? Who enforces contracts and settles disputes over property rights in the international area?

> Throughout history, powerful nations have tried to create international economic systems according to their own tastes and in harmony with their own interests. If the leaders of such great nations have thought about the subject at all, they have had no great difficulty in persuading themselves that the systems they were promoting also served the interests, if not the tastes, of the rest of mankind. In the past few decades, one power that has been pressing hard for the creation of an international system in its own image has been the United States.

The international economic system that prevailed roughly from the end of World War II until the beginning of the 1970's was characterized by the unprecedented prominence of international economic institutions and by a strong dose of hegemonial leadership by the United States, which not only placed its unmistakable stamp on international institutions but also enjoyed substantial leadership in the management of economic relationships.

But perhaps international economic relationships could be said to be characterized by a strong tendency toward competition, so that hegemonial leadership may be limited to a benign concern for enforcing contracts, correcting externalities and supplying public goods. The mid-ocean auctioneer and atomistic merchants could then fruitfully carry on their tasks.

The hypothesis that unorganized markets, with prices made by merchant intermediaries, had been the dominant market form throughout most of history may work fairly well, up until this century. But such markets surely have markedly declined. They have been largely replaced by fix-price markets, in which prices are set by the producers themselves (or by some authority); so they are not determined by supply and demand. It is of course granted that cost conditions, and sometimes also demand conditions, affect the prices that are fixed; but when these change, prices do not change automatically. Decisions, which are influenced by many other things than the simple demand-supply relationships, have to be made about them. That modern national and international markets are predominantly of the fix-price type hardly needs to be verified. It is verified by the most common observation.

Calls for a New International Economic Order (*NIEO*) have focused attention on the issues raised above, which can be summarized in the following questions: Who sets the rules of the game for international transactions and markets? Who has the power to initiate changes in such rules? Which international transactions are en-

*Professor of economics, Yale University. Edmar Bacha and Gerald K. Helleiner have been crucial partners in the preparation of this paper. Benjamin I. Cohen, Jorge Braga de Macedo, Gustav Ranis, Louka Papaefstratiou, and Ernesto Zedillo made helpful comments on an earlier draft.

couraged and which discouraged by the rules? How efficient and competitive are international markets?

The basic hypothesis is that the institutional framework within which international transactions take place has been historically rigged in favor of economic agents from the politically powerful countries, that is, on the whole rigged against economic agents from the less developed countries (*LDCs*). At least from a scholarly viewpoint, it is a virtue of the call for a *NIEO* that it tries to examine the nature of the whole system of international economic relations, besides raising more specific proposals for reform.

The debate over the *NIEO* has witnessed an unusual amount of sound and fury, including the cool fury of some economists who dismiss all *LDC* positions as the babbling of economic illiterates seized by a fit of passion. Particularly at the journalistic level, the picture is often drawn of an efficient, competitive, and liberal international order threatened by cartelizing or bureaucratizing pretensions of emotional, greedy, or ignorant *LDC* agents. This is why I have relied in this introductory section on extensive (hidden) quotes from the writings of four distinguished economists to state what I regard as the basic case for taking the call for a *NIEO* quite seriously, both at the academic and policy levels.[1]

What follows will discuss some of the *prima facie* inefficiencies and asymmetries of the international economic system, naturally stressing a few of special interest to *LDC*s. A modest pretension is to convince the reader to question the assumption that international markets and arrangements are as efficient, competitive and liberal as they can be. If such a view were accepted, a more fruitful dialogue between the North and the South could ensue.

[1]The first paragraph has been lifted directly from Arthur M. Okun, p.32. The third paragraph is also lifted completely from Raymond Vernon, p.12. The fourth paragraph paraphrases Marina v.N. Whitman, p.7. The sixth paragraph also paraphrases John R. Hicks, pp.x-xi. I beg the indulgence of these authors, especially of the last two.

I

It is a fair guess that to a Martian observer of our planet's economy, the most striking puzzle would be why a person growing cocoa in the tropics makes one-tenth of the wage of a man making aluminum ingots in cooler regions. After all possible explanations are given for this phenomenon, the suspicion remains that the world's labor force is not allocated in a Pareto optimal fashion. Large disparities seem to persist between different parts of the world in the returns to unskilled labor, much larger than disparities in the returns to capital, or to skilled labor.

The Martian observer may be told that the postwar liberal international economic order has encouraged a tendency toward a narrowing of the unskilled wage differential by promoting freer trade in commodities as well as freer capital movements. Beautiful Hecksher-Ohlin-Samuelson diagrams will help the Martian understand how a freer movement of goods and capital work toward such a purpose. Indeed, it may be suggested that insofar as wage gaps persist between the tropics and the cool regions this could be due to foolish tropical barriers to commodity trade or to inflowing capital.

Being naive, our Martian may ask: Would it not be simpler to allow unskilled labor to move from where its marginal product appears low to where it seems to be high, thus making everybody potentially better off? After all, during the nineteenth century there were massive and persistent movements of unskilled white labor from Europe to "regions of recent settlement." A possible answer to this query is to tell the Martian that he is being impractical and that he should go back to where he came from. Alternative answers could express concern about an apparently unconquerable tropical incontinence, or to complications arising from the existence of public property.

Given the ingenuity of our profession, it is conceivable that models could be built in which gaps in wages between the tropics and cooler regions would be compatible

with an optimal allocation of world resources. What is remarkable is how quietly our profession, otherwise so intolerant of bureaucratic limitations on the freedom of economic agents, accepts and takes for granted governmental barriers to the free flow of unskilled labor across artificial national boundaries. An economist working for the *U.S.* government, for example, may have to simultaneously argue for the desirability of nondiscriminatory or national treatment for *U.S.* direct foreign investment in Mexico, for the importance of freer emigration of skilled labor from Eastern Europe and for the necessity for the Mexican government to force its unskilled masses to stay on their side of the river. Matters may even be more complicated for that economist's professional conscience: he or she may have to urge Mexicans to export oil and gas but to go easy on their exports of steel. A comparison of the legislation and practice of industrial countries regarding immigration with that regulating their merchandise imports reveals that they mercantilistically prefer their commodity imports raw and their immigrants polished.

The relative prices and income distribution which would exist in a world of no interference with the free flow of unskilled labor are very likely to be different from those which were generated under the international economic system of the last thirty years. A plausible conjecture is that this counterfactual situation would have been more favorable to many economic agents in *LDC*s. Hence it is not surprising that some *LDC* spokesmen regard the international economic system as rigged against their interest, and tending to keep the value of tropical labor below that of cooler regions' labor. They may attempt using countervailing power to offset Northern monopsony power. Without *laissez-passer* for unskilled labor, the case for *laissez-faire* is weakened.

II

One group of economists believes that all prices quickly approach marginal social costs regardless of apparent impurities in real world markets, whether national or international, and in spite of public or private efforts to thwart what they regard as spontaneous economic forces. Another group prefers to take more seriously the apparent departures from the assumptions necessary for generating purely competitive results, and to explore the consequences of oligopoly, oligopsony, and quasi rents. Those in the first group, for example, regard the *OPEC* or the diamond and nickel cartels as transient phenomena, unimportant in the long run. Such departures from competition are doomed to failure, they argue, and presumably the quasi rents they capture in the short run will not influence significantly long-run values or the distribution of income and wealth. Economists of that persuasion will find little interest in what follows.

Just like the farmers and the miners from the Midwest and the West of the United States late in the last century, *LDC*s feel that the markets for what they sell and what they buy are manipulated by economic agents they do not control, and who tilt market results against them. The populist suspicions about middlemen, the railroads, the banks, and of remote concentrations of economic power in general, are echoed in the calls for a *NIEO*, which therefore may be said to be as much of a "cooperativism of the poor" as a "trade unionism of the poor." The same issues, of course, reappear within *LDC*s, between town and country, and between the informal and the organized sectors.

Crucial complexities in international markets for tropical products are far from fully captured by standard demand and supply schedules and assumptions about the instantaneous clearing of markets. Since last century it has been observed that while production of those commodities was often in the hands of not far from atomistic *LDC* producers, the marketing, storing, grading, and processing was handled by non-*LDC* economic agents of nonatomistic dimensions with privileged access to credit, and who carefully controlled market information. For example, in 1888 the association

of Cuban sugar producers had to set up a system of daily telegraphic reports from New York and London to learn not only the commodity prices, but to receive estimates of Cuban sugar production. It is said that all such information was controlled by the Willett and Gray firm, a member of the *U.S.* sugar trust, was sent by Western Union, and was distributed in Cuba by the Associated Press.[2] More recently, the U.S. Justice Department has alleged that the New York Coffee and Sugar Exchange, Inc., which many thought a reasonable example of an auction or flexprice market, and various coconspirators were artificially influencing sugar prices, while an E.E.C. commissioner charged that a few companies had indulged in "grave and scandalous speculation" in sugar markets.[3] One may entertain a reasonable doubt as to whether it is correct to interpret *LDC* efforts to extend an international sugar agreement as nothing more than the replacement of efficient competitive markets by a cartel run by *UNCTAD* bureaucrats. One may also add that sugar is a product for which the presumed concern of the cooler countries for a rational allocation of resources, free from artificial distortions, shines not too brightly in historical perspective. More generally, with agricultural sectors in industrialized countries so plagued with departures from *laissez-faire*, one marvels at their fervent advocacy of *laissez-faire* in international agricultural markets.

In the case of *LDC* mineral exports, production as well as marketing and processing has been controlled by non-*LDC* economic agents, of a size only the hopelessly myopic could call atomistic. Vertically integrated, transnational corporations internalized most markets between mines and the consumer of finished products, leaving the auctioneer in mid-ocean to amuse himself with thin, residual and unstable markets. It is bizarre to hear warnings of how Jamaica is cartelizing the bauxite market from economists who hardly mention the organization of the aluminum industry nor the ghostly nature of "the bauxite market," more the creation of aluminum company accountants than the domain of auctioneers.

It may be useful to conceptualize many historical international mineral markets as having been organized by a few transnational corporations, which ran species of commodity stabilization schemes, where unexpected changes in economic conditions were reflected not on price movements but first of all in changes in the levels of the buffer stocks controlled by those corporations and/or in changes in the speed with which different types of customers were serviced. Besides subtle "customers' relationships" those corporations have maintained links with governments, particularly those of parent and host countries. Corporate central planning boards decided on investment projects on the basis of long-term forecasts, rather than just on the basis of present market conditions (futures' markets were also internalized). It can be argued that such arrangements were often economically superior to anything the mid-ocean auctioneer and a mob of atomistic economic agents could have wrought. This may be so, but to analyze such situations standard demand and supply schedules provide limited insights.

The blunt fact is that scientific knowledge about the operation of international commodity markets is very scanty. The hypotheses that these markets function as if they were competitive and with desirable properties such as informational efficiency have been seldom put to rigorous test. Sheer repetition of hypotheses should not be confused with established fact, and helps little when making the difficult choice between imperfect buffer stock arrangements and imperfect unregulated markets.

III

It may be a good thing that there are economists with whom one has to argue at some length the thesis that international markets are far from competitive, that the

[2]See Manuel Moreno-Fraginals, p.20.

[3]See the *Wall Street Journal*, p.2; and *The (London) Economist*, p.66.

international economy has room for improving both its efficiency and its equity, and that political power can be translated into favorable rules of the game and the privilege of initiating changes in those rules. Space allows only a few more examples of such propositions, which many will find bland.

The macroeconomic management of the capitalist world economy has been in the hands of representatives of a few industrialized countries. Other nations, particularly the *LDC*s, have been forced to be passive spectators, even though world economic conditions can influence their welfare significantly. The record of the macroeconomic managers over the last few years is not a spectacular one. The massive unemployment and idle capacity of the industrialized countries are not just monumental wastes in themselves, but by directly and indirectly discouraging international specialization they also impose waste on the rest of the world. Surely issues such as the smooth servicing of *LDC* debt cannot be discussed in isolation from those of world macroeconomic management and outbreaks of protectionism in industrialized countries. At least some *LDC*s must increasingly participate in world macroeconomic management.

Advocates of the flexibility and resourcefulness of decentralized international market can point with pride to the mushrooming of international private lending to *LDC*s over the last ten years or so. But a curious thing seems to be happening here. Orthodox voices are being increasingly heard arguing that this stronghold of *laissez-faire* is in need of regulation. International lenders, it appears, compete too much and a paternalistic International Monetary Fund may be needed to insure "orderliness" in that market. A credit cartel to such orthodox observers may be required to correct market failures (of an unspecified sort); the same observers would undoubtedly scorn proposals to regulate direct foreign investment by transnational enterprises. Finally, it may be interesting trying to explain to a Martian observer of the world economy the rationale for the continuing *U.S.* trade embargo against countries such as Cuba and Vietnam.

However it is time to turn to an obvious question: Is it that the *NIEO* seeks the establishment of international markets resembling those of a neoclassical textbook? The answer, of course, is no. Several of the *LDC*s which have spearheaded the drive for a *NIEO*, and the public and private economic agents behind that drive may best be conceptualized as new oligopolists, trying to break into world markets dominated by old oligopolists. The new oligopolists want to exercise a greater share of market power, whether in the markets for their raw material and primary product exports, or in those for their new manufactured exports, or in those for their imports of machinery and technology. The new oligopolists will set up their own transnational corporations for this purpose, or will try to manipulate existing ones. The incentive is not just a cut in declared oligopolistic profits and rents, but also a share in the "perks" attached to control of hierarchical bureaucratic organizations, which under standard accounting conventions are recorded as business costs.

While this may not be their intention, the eruption of new oligopolists into world markets could in fact lead directly and indirectly to greater competition and a closer approximation to textbook ideals, at least in some markets, and for a while (for example, copper). Populist agitation, economic historians remind us, helped to bring about antitrust legislation, a more rational control over money in the United States and contributed to ambiguous forms of countervailing power and to regulations stabilizing oligopolistic market situations. In short, the call for a *NIEO* may be interpreted partly as a call for adjusting to "two, three, more Japans" within the world capitalist economy. The *LDC*s such as Algeria, Brazil, and the Philippines, with a growing industrial might and high capacities to absorb capital and technology, are interested neither in wrecking international markets nor in shutting themselves off from them; they seek first to gain a greater share of the action in those markets,

and then to participate in "organizing" trade in them. Wise old oligopolists will understand such motivation, even as their anxieties increase over their possible loss of industrial and technical hegemony.

Where does that leave the poorest *LDC*s, or the poorest groups within most *LDC*s? One should not rule out the possibilities that in some cases those groups may benefit from the *NIEO*. The new oligopolists need political legitimation and votes at the United Nations. A given policy proposal can benefit countries with very different social systems, leading to startling alliances, such as the close Brazilian-Cuban cooperation in the negotiations for an international sugar agreement. But even a total acceptance of *NIEO* proposals, taken by itself, is unlikely to significantly improve in the short run the welfare of the poorest half of *LDC* families, nor to significantly worsen, one may add, the welfare of the poorest families in the North. While the economist should be skeptical of claims by some of the new would-be oligopolists to represent and work for the poor, he or she should also not take as self-evident the pretension of old oligopolists to be helpless minions of The Market Force, nor the embodiment of The National Interest.

During the last one hundred years latecomers to industrialization, such as Germany, Italy, and Japan, were not smoothly integrated into the world economy. Even today, deep suspicions remain in Europe and the United States as to whether Japan plays fair, a suspicion which is no doubt reciprocated. For the sake of world peace one hopes that the process of adjusting to late-latecomers will be less painful.

REFERENCES

John R. Hicks, *Economic Perspectives; Further Essays on Money and Growth*, Oxford 1977.

M. Moreno-Fraginals, "Cuban-American Relations and the Sugar Trade," mimeo., Oct. 1977.

A. M. Okun, "Further Thoughts on Equality and Efficiency," in Colin D. Campbell, ed., *Income Redistribution*, Washington 1977.

Raymond Vernon, *Storm Over the Multinationals; The Real Issues*, Cambridge, Mass. 1977.

M. v.N. Whitman, "Sustaining the International Economic System: Issues for U.S. Policy," in *Princeton Essays in International Finance*, No. 121, June 1977.

The (London) Economist, Sept. 24, 1977.

Wall Street Journal, Oct. 18, 1977.

Alternative Trade Strategies and Employment in *LDC*s

By ANNE O. KRUEGER*

The 1970's have witnessed significant changes in the trade policies and strategies of many *LDC*s. In the 1950's and 1960's most of them adhered to policies of import substitution behind highly restrictive quantitative controls, intensified by overvaluation of the exchange rate with attendant disincentives for export. In the past decade, there has been a marked reduction in the degree of bias toward import substitution. Even in countries where quantitative restrictions and tariffs continue to provide inducements for production for the domestic market much greater than incentive for sale abroad, bias is less extreme than in the past. In other countries, notably Brazil and South Korea, bias has been completely reversed, to a point where one might even claim a bias towards the foreign market and against the home market.

This shift in trade policies has resulted from a number of factors, some specific to individual countries, but chiefly because of evidence that the excesses of import substitution were detrimental to growth. The fact that there have been policy switches has enabled economists to attempt to estimate the differences in growth prospects for countries employing alternative trade strategies. Interestingly, the several studies that have been made (Constantine Michalopoulos and Keith Jay; Michael Michaely; Bela Balassa; the author, 1978) using widely different methodologies have nonetheless reached quantitatively similar conclusions as to the increase in growth rates resulting from a switch.

Although trade policies have altered, other aspects of growth performance in *LDC*s have been increasingly questioned. Important among these is the growth of employment opportunities. It has become fashionable, especially among policymakers, to conclude that achieving a high rate of growth of *GNP* will not necessarily entail the growth of opportunity for new entrants to the labor force. Thus, the demonstration that export-oriented trade strategies results in superior growth performance may have come too late: pessimists may conclude that faster growth does not necessarily imply more employment.

In light of this view a natural question, but one that has not previously been examined systematically, is the relationship between alternative trade strategies and employment. It is possible that the observed unsatisfactory growth of employment resulted largely from the choice of an import substitution strategy, or at least that export promotion is more compatible with employment growth than is import substitution. To be sure, until recently one would have simply asserted that faster growth associated with one strategy would in itself lead to a more rapid upward shift in the demand for labor, and in all probability, the growth effects of alternative trade strategies dominate the compositional and substitution effects originating from those alternatives. However, given the prevailing skepticism about the effects of increased rates of *GNP* growth on employment and income distribution, it seems worthwhile investigating the issue in some depth.

The link between trade strategies and employment is a complex one. It involves not only trade strategies themselves and the structure of labor markets, but also entails

*Professor of economics, University of Minnesota, and senior research staff, National Bureau of Economic Research. The paper was written while I was visiting fellow at the Australian National University, and benefited from comments made during presentation of a seminar there, and also at the Development Research Centre, International Bank for Reconstruction and Development. The National Bureau's project on Alternative Trade Strategies and Employment is funded in major part by the Agency of International Development whose support is gratefully acknowledged.

examination of the various factors which contribute to distorting goods and factor markets in *LDC*s. In an effort to sort out some of the issues involved in the trade strategies-employment relationship and also to obtain estimates of the quantitative significance of the links between trade strategies and employment, the National Bureau of Economic Research (*NBER*) has been conducting a research project on the topic. A major part of the *NBER* project has consisted of individual country studies, now nearing completion, undertaken by authors analyzing the situation in their particular countries. This paper represents a preliminary report on some of the findings that have emerged to date.

There are two aspects of the project which require discussion. First, there is the underlying analysis of the link between trade strategies and employment. Second, enough empirical results are available from the country studies to provide some data on some aspects of the employment-trade relationship. The countries studied and the authors are Brazil: Jose Carvalho and Claudio Haddad; Chile: Vittorio Corbo and Patricio Meller; Colombia: Francisco Thoumi; India: V.R. Panchamukhi and T.N. Srinivasan; Indonesia: Mark Pitt; Ivory Coast: Terry Monson and Jacques Pegatienan; Kenya: Peter Hopcraft and Leopold Mureithi; Korea: Wontack Hong; Pakistan: Stephen Guisinger; Thailand: Narongchai Akrasanee; Tunisia: Mustapha Nabli; Uruguay: Alberto Bension. Within a common framework of analysis and empirical methodology, all authors have attempted to estimate the trade-employment relationship for their particular country.

Before turning to the analysis of the link between trade strategies and employment, one preliminary matter must be cleared up. That pertains to the determinants of employment. After all, in a neoclassical economy, any shift in the demand curve for labor is reflected in a change in the real wage and employment changes only to the extent that the labor supply is responsive to the real wage. More generally, depending on the underlying structure of the labor market, an upward shift in the demand for labor can result in almost anything: in a Harris-Todaro world, it can even result in increased unemployment. In a world in which unions or governments are sufficiently powerful, an upward shift in the demand for labor may result in an increase in the real wage, regardless of the underlying employment situation. Thus the ways in which employment—in total and in the urban sector—responds to changes in demand for labor are functions of many variables.

For purposes of analyzing trade strategies, focus is put upon determinants of the demand for labor. The word employment is used synonymously for "demand for labor." This means that such phenomena as differences in the choice of technique or industry as a consequence of different trade strategies are evaluated in the project. However, the factors determining how much of an upward shift in the demand for labor results in greater employment and how much in higher real wages are considered to lie outside the scope of the research (see the author, 1977).

Turning to the trade strategies-employment relationship, there are three levels of linkages with simultaneity among them. First, there is the effect of the choice of trade strategy on the overall rate of growth, and therefore on the rate of growth of employment opportunities. As already indicated, experience strongly suggests that there is such a link, but it is not dealt with in the present project.

Second, there is the effect on the demand for labor *via* the influence of the trade strategy on the composition of output. If one trade strategy results in a higher proportion of *GNP* originating in labor-itensive industries, the selection of that strategy will unequivocally result in a higher demand for labor (at a given real wage) than will the selection of the alternative. It should be noted here as elsewhere that it is not only the choice of trade strategy that can matter: the degree to which the strategy is pursued will affect the composition of output. Korea's export industries, for example, form a

larger fraction of her output (and therefore are more heavily weighted in determining the demand for labor) than if the emphasis upon export promotion were somewhat less.

Third, there is the effect that the trade regime has on factor prices: to the extent that, for example, import substitution results in incentives for use of capital-intensive techniques (due to overvaluation of the currency, and perhaps also to policies that discriminate in favor of applications for foreign exchange for financing imports of capital goods), it may affect the demand for labor. To be sure, such an effect must be sectoral rather than total, unless one wishes to argue that the total capital stock is different under that strategy than it would have been under the alternative.

The *NBER* project is focused upon the second and third links between trade strategies and employment. Because of space limitations this progress report is devoted simply to the second link: the different factor proportions employed in industries with differing trade orientations. At an analytical level, the underlying theory can be taken directly from the Heckscher-Ohlin-Samuelson (*HOS*) trade model, and then transformed into empirically testable hypotheses about factor proportions in importables and exportables. In practice, however, it proves useful to go beyond the usual 2 × 2 × 2 framework, and to consider the predictions of an *HOS*-like model with three factors (the third presumably being a raw material), many commodities, and many countries of the type I have previously discussed (1977b).

Within that framework, commodities are partitioned, in accordance with their characteristics, into natural resource-based and *HOS* goods, and, within each of those categories, into exportables, import-competing, and noncompeting-import categories according to trade flows.[1] Where relevant, import and export data are also broken down by origin and destination, respectively. The multicountry, multicommodity trade model immediately suggests the hypothesis that the commodity composition and factor proportions of tradables with countries with significantly different factor endowments may be quite different than that with countries with similar endowments. To illustrate, one would expect that Brazil might have a comparative advantage among *HOS* goods in more labor-intensive commodities in her trade with industrialized countries than in her trade with her Latin American Common Market trading partners.

Table 1 presents some of the preliminary findings as to the labor content of trade per unit of domestic value-added (*DVA*) in domestic production of tradables. The "direct" column presents the ratio of the

[1]Noncompeting imports are those commodities which are imported for which domestic production cannot substitute within the relevant price range.

TABLE 1—RATIO OF LABOR COEFFICIENTS IN EXPORTABLES TO THOSE IN IMPORT-COMPETING INDUSTRIES[a]

Country and Category	Direct Requirements	Direct plus Home Goods Indirect
Brazil[b] all trade	1.07	2.67
Chile all trade	.81	1.07
Colombia all trade	1.93	-
Indonesia		
All trade except oil	2.09	1.92
Trade excluding all raw materials	1.73	1.58
Ivory Coast *HOS* goods	1.38	1.35
Kenya all trade	.72	-
Korea all trade 1966	1.25	1.23
All trade 1973	.79	.96
Pakistan *HOS* goods	1.41	-
Thailand all trade	2.21	1.70
Tunisia *HOS* goods[c]	1.32	1.22
Uruguay *HOS* trade with developed countries	1.87	-

Sources: Individual country study manuscripts.

[a]Estimates are for different years between 1966 and 1973 depending on data availability in individual countries.

[b]Brazilian data are per unit of output rather than per unit of value-added, and therefore are not comparable with the ratios for other countries. Indirect includes agricultural indirect inputs.

[c]Unskilled labor only.

number of man-years of labor direct requirements per *DVA* in exportables to the number in import-competing industries, while the "direct plus indirect" refers to the so-called Corden measure—direct requirements, plus indirect requirements in home goods. It does not include indirect requirements of tradables, although those numbers were also computed by most country authors and do not significantly alter the results.

As can be seen from Table 1, in nearly all cases and in all cases of *HOS* goods, exports are less labor intensive than import-competing industries. Indeed, the differential in labor requirements between export and import-competing industries is quite notable in many countries—virtually 2:1 in Columbia, Indonesia, Thailand, and Uruguay. Moreover, as discussed below, Keyna's and Chile's ratios reflect a high proportion of trade with *LDC*s. In light of the Leontief Paradox findings, and also in view of the alleged importance of factor market distortions in many of the *LDC*s, it is perhaps surprising that the greater labor intensity of exports shows up so consistently across countries, even though the data pertain chiefly to all trade. The differentials in labor intensity are all the more remarkable when it is recognized that there are usually incentives for capital-intensive production of all commodities, including exportables. As such, it is likely that the "optimal" labor intensity of most activities is even greater than that indicated by the data included in the table.

The fact that exportables are more labor-intensive than import-competing industries conforms to the *HOS* model of international trade. Perhaps a somewhat more unexpected result, shown in Table 2, pertains to the differentials in labor intensity between exports to developed countries and those to other developing countries. With the exception of Thailand, all the countries for which there was enough trade to warrant the separate calculation of labor coefficients were found to have remarkably different products, and therefore factor proportions, in their trade with developed countries than with other developing countries. Uruguay, for example, exports products to developing countries that are less labor using than even the production competing with imports from developed countries. Likewise, Chile's exports to developed countries require almost 40 percent more labor per *DVA* than do her total exports.

TABLE 2—RATIO OF LABOR REQUIREMENTS FOR EXPORTS TO DEVELOPED COUNTRIES TO THOSE FOR *LDC* TRADE

Country	Direct	Direct plus Home Goods Indirect
Chile[a]	1.40	1.31
Kenya	1.36	–
Thailand	1.10	1.03
Uruguay	1.84	–

[a] For Chile, the ratio is labor per *DVA* of exports to developed countries divided by labor per *DVA* for total trade.

A natural interpretation of this phenomenon is that the prospects for trade among the *LDC*s are indeed limited, as customs union theory forecasts. In fact, the large disparity in factor proportions between the various destinations strongly suggests that it is the high-cost import substitution industries which find their only export outlet in other developing countries.[2] Insofar as developing countries are relatively abundantly endowed with unskilled labor and relatively short of capital, trade with other *LDC*s is likely to increase the imbalance in factor availability, whereas trade with the developed countries may serve as means of exchanging abundant factors for scarce.

Another finding of considerable interest that emerges from a number of the studies is that the skill requirements in exportables

[2] Anyone familiar with the trade regimes and protectionist policies of *LDC*s will be aware that most of them erect high trade barriers against any goods which are domestically produced. To the extent that such barriers cut off any trade which would follow from "natural" comparative advantage, it would intensify the capital intensity of one *LDC*'s exports to another *LDC*.

TABLE 3—RATIO OF DIRECT SKILL COEFFICIENTS PER *DVA* IN EXPORTABLES TO IMPORT-COMPETING INDUSTRIES

Country	Ratio
Brazil, all industry	.98
Chile, all trade	.83
Indonesia, all trade	.81
Ivory Coast, *HOS* only	.99
Kenya	.78
Tunisia	.62

and import-competing production are virtually as disparate as the labor inputs. Data are given in Table 3. For example, Chile's exports appear to require about 83 percent as many skill units as her production of import-competing commodities. In general, exportables appear to have lower requirements of skilled man-years per *DVA*, regardless of the measure of skills used. The results appear to be fairly robust both with regard to commodity classifications and also with regard to the definition of skills. This finding lends support to the notion already prevalent elsewhere that capital-intensive industries are also industries where skill requirements are high, whereas unskilled labor-intensive industries appear to have both lower capital requirements and lower skill requirements per *DVA*.

Thus, the findings both with regard to skills or human capital, and with regard to unskilled labor, seem to reinforce the notions set forth in the neoclassical model of international trade. The Leontief Paradox results, as well as the theory underlying the analysis of trade in the presence of factor market distortions, have provided some ground for skepticism that comparative advantage in labor-abundant countries might lie in production of labor-intensive goods (as contrasted with the surely correct notion that comparative advantage lies in producing whatever one produces with relatively labor-intensive techniques). Nonetheless, the data from the countries covered by the *NBER* project indicate that, once allowance is made for direction of trade, the labor-abundant developing countries probably would be well-advised to specialize in the export of labor-intensive products.

A paper of this length cannot possibly do justice to the scope and range of results that are emerging from the individual country studies and the other outputs of the *NBER* project. There will, however, be a volume containing the salient findings from the individual country studies which should provide considerably greater information to interested parties. What is already clear is that the findings of the country studies support the view that altering trade strategies toward a greater export orientation will certainly be consistent with the objective of finding more employment opportunities: skepticism based on Leontief Paradox or factor-market distortion considerations does not seem to be warranted.

REFERENCES

B. Balassa, "Exports and Economic Growth: Some Further Evidence," mimeo., Washington, July 1977.

M. Michaely, "Exports and Growth: An Empirical Investigation," *J. Develop. Econ.*, Mar. 1977, *4*.

C. Michalopoulos and K. Jay, "Growth of Exports and Income in the Developing World: A Neoclassical View," disc. paper no. 28, Agency Int. Develop., Washington 1973.

Anne O. Krueger, *Foreign Trade Regimes and Economic Development: Liberalization Attempts and Consequences*, Cambridge, Mass. 1978.

———, (1977a) "Alternative Trade Strategies, Growth, and Employment," in N. Akrasanee et al., eds., *Trade and Employment in Asia and the Pacific*, Honolulu 1977.

———, (1977b) *Growth, Distortions, and Patterns of Trade Among Many Countries*, in *Princeton Studies in International Finance*, No. 40, 1977.

Some Aspects of Technology Transfer and Direct Foreign Investment

By RONALD FINDLAY*

The creation and diffusion of new technology is undoubtedly the major determinant of economic growth. Historians such as Carlo Cipolla, David Landes, and Nathan Rosenberg, among many others, have given extensive descriptions and insights into this process for the more recent centuries. In economic theory the role of technology has, however, largely been as an exogenous factor which the economist regards as part of his data. This seems to be true of all schools. Neoclassical economists take their smoothly differentiable production functions as given, just as Joan Robinson's "book of blueprints" is apparently distributed free of charge by some benign technocratic agency. The economists' separation of factor substitution within a given set of techniques from technological change in the sense of the addition of new techniques is an increasing source of uneasiness and is perhaps a needless obstacle to understanding. The major problem is that the availability of techniques of production is, to some extent at least, a function of the resources devoted to finding them. The textbook diagrams of a given map of isoquants to represent a menu of alternative technical possibilities from which a selection is made on the basis of the scale of output desired and relative factor prices are not very helpful when the question of whether or not a particular technique is on the menu depends on scale of output and relative factor prices. A clear statement of this interdependence between market factors and technology and the reasons why it causes problems in the conventional economist's conceptualization of technological change and diffusion is given by Rosenberg (ch. 4). In this paper I shall attempt to build a very simple model of technology transfer that explicitly incorporates the interdependence between output, costs, and technology that Rosenberg stresses.

The problem of technology transfer is one that has aroused much controversy. On the one hand it is alleged that multinational corporations, which are the repositories of much of the more sophisticated technology, are unduly restrictive in the transmission of this technology to the developing countries in which they operate and charge excessively for whatever is transmitted. On the other hand, it is frequently asserted that the technology that is transmitted to developing countries is not "appropriate" to their relative factor proportions. The emphasis in the problem of choice of techniques has shifted from making the socially optimal selection from a given set of techniques to one of generating the techniques themselves. However the new literature on appropriate "intermediate" or "progressive" technologies, such as that of Keith Marsden, has so far been of a somewhat utopian nature. The model that follows is an attempt to link the discussion surrounding the broad questions about appropriate technology transfer with elementary micro-economic analysis of factor substitution and innovation.

In Figure 1, suppose initially that the curve QQ' represents the unit isoquant of a constant return to scale production function of conventional neoclassical theory. Let point a represent the technique of production used in advanced countries. Relative factor prices are such that cost is minimized by the use of this technique. If production of the commodity is to be introduced into a less developed economy, the relative price of labor, indicated by the slope of the line FF', will be much lower. The appropriate technique of production should now be represented by the point b, if the conventional isoquant were to exist.

*Columbia University.

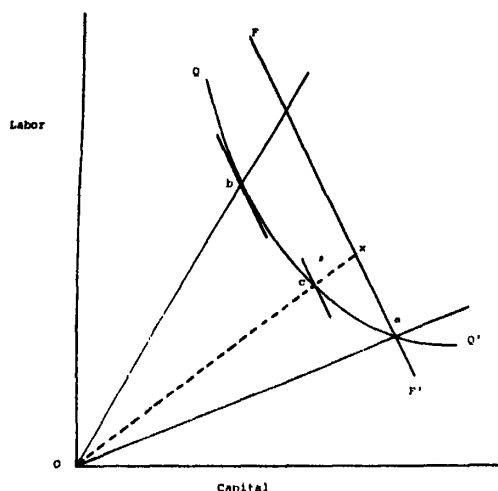



FIGURE 1

The question now arises as to whether the alternative techniques represented by the conventional isoquant depict existing specific processes that can be introduced directly into factories or are instead merely indications of what can be achieved within the existing framework of general scientific knowledge. There are problems with either conception, as was pointed out in some acute observations by W.E.G. Salter (ch. 2) in his pioneering study. In the case of the first alternative, the separation of technology from market forces that is an essential feature of conventional production theory goes by the board and, furthermore, the range of techniques available would be restricted to those that are currently commercially viable. The problem with the second alternative of including all possible designs is that one escapes from the contamination of technology by market forces at the price of being left with an extremely nebulous and empty concept. It may be asked how one knows what are the boundaries of technical choice constrained only by general scientific knowledge. It is certainly not possible to say that the less developed economy should choose technique *b* when it may require considerable expenditure to actually be made operational. Salter himself prefers the latter alternative and, long before the subject of appropriate technology became fashionable, made the following wise observation:

> One wonders how far the use of highly mechanized techniques in such (less developed) economies is due simply to the fact that the details of western technology reflect the need to save labor, and that this set of detailed designs may be quite inapplicable to economies where labor is cheap. Perhaps underdeveloped countries should use western knowledge to develop a detailed technology aimed at methods that are modern and technically efficient but not mechanized. [p.15]

I agree with Rosenberg when he criticizes Salter's choice of the second alternative. He states that from an economic point of view there is no real difference between an innovation in production being due to an increment in engineering or scientific knowledge. As he says, "It is not, after all, scientific knowledge which is ultimately valued in the economic sphere but, rather, knowledge in a form which is directly appliable to productive activity" (p.65).

If we follow Rosenberg, however, we confront the problem of relative factor prices and the volume of production influencing the availability of techniques. We seem to be left with the choice for the less developed economy between technique *a* which is feasible but not desirable, and technique *b* which would be desirable but is not feasible in the absence of allocation of resources to its determination. Moreover if resource allocation to the development of new technology is to be considered, then would it still follow that *b* is the appropriate technique?

I shall now try to develop the simplest model in which the availability of techniques is an endogenous variable. We start from the idea that a new process which does not lower costs of process *a*, evaluated at factor prices of the developing country, would not be worth any discovery expenditures. In Figure 1, consider any ray from the origin such as Ox, with the slope

steeper than Oa. A new process corresponding to point x on the line FF' would not reduce the factor cost of process a at all and so would not be worth any expenditure on discovery. However, suppose that incurring further expenditure on design enables us to reduce factor cost by reducing both labor and capital inputs proportionately along Ox. Then we should continue this expenditure until we reach a point along Ox at which the marginal increment in expenditure on the design is equal to the marginal gain in the capitalized value of the stream of savings in factor costs that it brings about over the life of the project. Suppose that this point on Ox is represented by c, the intersection of Ox with the curve QQ'.

For each ray from the origin we can imagine the experiment being repeated. Let the curve QQ' represent the locus of points that, similar to c, are optimal with respect to the balancing of costs and benefits on new techniques when the choice of technique is confined to a particular labor intensity. Notice that QQ' is no longer a conventional isoquant. Each point on it is *not* independent of relative factor prices and the volume of output expected to be sold at the corresponding factor cost. The information needed to generate what may be called the "potential isoquant" QQ' is the fixed cost of design for each technique, the relative factor prices FF' and the demand curve for the product to enable the total flow of savings in factor cost to be computed.

Each point on QQ' represents a particular labor intensity with an associated fixed design cost and a present value of benefits from factor-cost reductions. It is readily seen that these benefits are a concave function of labor intensity since they are zero for the technique Oa, rise to a maximum at Ob, and decline thereafter. I hypothesize that design costs are a monotonically increasing function of the labor intensity of techniques. This seems to be intuitively plausible since the more capital-intensive techniques are closer to the already operating techniques in advanced countries. In his valuable empirical study, David Teece states that ". . . the relevant skills for highly capital-intensive industries, such as chemicals and petroleum refining, are more easily transferred internationally than are the requisite skills in the machinery category" (p.259). While his empirical evidence refers to differences in capital intensity between industries, he also believes it to be true about differences between processes in the same industry. He says, "This is consistent with speculation that international transfer is no more difficult than domestic transfer when the underlying technology is highly capital intensive. The perceived reluctance of multinational firms to adapt technology to suit the capital-labour endowments of less developed countries could well be rooted in the desire to avoid escalating transfer costs to unacceptable levels" (p.259).

In Figure 2, the optimum process is determined, showing the benefits resulting from factor-cost reductions as a function of the labor intensity of the new processes as the concave curve BC. The design and adaptation costs of the new processes are shown as a function of the labor intensity in the convex curve GH. The optimal labor intensity Ox^* is where the distance between BC and GH is at a maximum. In other

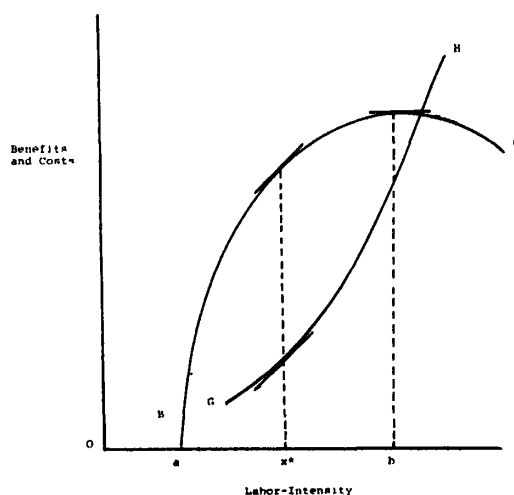



FIGURE 2

words, Ox^* is the labor intensity at which the marginal increase in social benefits from reductions in factor costs is just equal to the marginal increment in design and adaptation costs. We therefore see that the true optimal process is not *Ob* which maximizes the total benefit from factor cost reductions since this is too expensive to achieve. On the other hand, it is more labor intensive than the technique *Oa* already in use in the advanced countries, since this does not economize sufficiently on relatively scarce capital.

The reader will, of course, be aware that my argument has made several simplifying assumptions. One is that conditions such as the relative factor prices and demand for the product remain stationary over time. Another is that I assume continuous variation in techniques instead of a discrete number. These assumptions however may be relaxed in a fuller and more formal analysis without affecting the spirit of the approach. One advantage of the method of presentation adopted is to show how the conventional formal apparatus of neoclassical production theory in terms of smooth isoquants and so on can carry over into this very different context. I therefore do not agree with Rosenberg when he states that this approach is "distinctly less helpful" when the problem of technological change is studied. We have seen how the isoquant concept can be modified to take account of the very interdependence between relative factor prices and research expenditure on developing new techniques that Rosenberg rightly considers so vital.

My approach is similar in spirit to that proposed by Anthony Atkinson and Joseph Stiglitz who stress the importance of what they call "localized" technical progress as opposed to entire shifts in the production function. My analysis proceeds in two stages of looking at the cost and benefits of a sequence of "localized" technical improvements and then choosing the optimal one from among them. A related contribution is that of Kenneth Arrow (ch. 6). He assumes a single improvement and compares the incentive to make that improvement under monopoly and perfect competition, finding it to be higher in the latter case. A similar result could be obtained from my model since the benefits from any particular technical improvement depend upon the volume of production, which with a given demand curve will be lower under monopoly.

This analysis shows that for the same relative factor prices the gain from introducing new techniques is higher the larger the volume of demand. This would indicate that countries such as Brazil and India are likelier to generate "intermediate" technology than smaller economies with the same degree of capital scarcity. One might therefore expect that there will be profitable opportunities for international sale of technical knowledge between countries at the same stage of development as measured by capital-labor ratios but differing in scale. The Heckscher-Ohlin and Linder hypotheses combine in this connection. The form which this trade takes might be sale of patents and payment of royalties or direct foreign investment by larger less developed economies in smaller ones, for which there appears to be some evidence already.

Space does not permit an extensive discussion of more dynamic aspects of technology transfer and foreign investment. Instead, I shall briefly summarize an approach adopted in my forthcoming paper which combines the view associated with Thorstein Veblen and Alexander Gerschenkron that the greater the relative backwardness of a country the faster the rate at which it can catch up, and the role of direct foreign investment in stimulating technical advance in developing economies. The rate of technical change in the developing region is made an increasing function of the gap between technology levels in itself and an advanced region and of the proportion of foreign to domestic capital within the backward region. With the rate of technical progress in the advanced region given exogenously, a formal dynamic model is

constructed in which there is a steady state with all variables growing at this exogenous rate and certain equilibrium ratios of technology levels in the two regions and of foreign to domestic capital. As historians such as Cipolla, Landes, and Rosenberg, and also Arrow (ch. 7) have stressed, the diffusion of innovations has always proceeded through direct contacts between the possessors and adopters of new technology. The role of migration in earlier times is now largely carried on through the aegis of the multinational corporations. In this model, the rate of growth of foreign capital is an increasing function of its after-tax rate of profit and the rate of growth of domestic capital is an increasing function of the domestic profit rate and the tax proceeds on the profits earned by foreign capital. An increase in the tax rate on foreign capital reduces the equilibrium ratio of foreign to domestic capital but also widens the relative technology gap between the advanced and backward regions. Foreign and domestic capital are regarded as distinct factors of production in this model, reflecting the fact that capital, management, and technology are inextricably combined, as pointed out in the pioneering work of the late Stephen Hymer.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk-Bearing*, Amsterdam 1974.

A. B. Atkinson and J. E. Stiglitz, "A New View of Technological Change," *Econ. J.*, Sept. 1969, *79*, 573–78.

R. Findlay, "Relative Backwardness, Direct Foreign Investment and the Transfer of Technology: A Simple Dynamic Model," *Quart. J. Econ.*, forthcoming.

S. Hymer, "The International Operations of National Firms: A Study of Direct Investment," unpublished doctoral dissertation, Mass. Instit. Technology 1960.

K. Marsden, "Progressive Technologies for Developing Countries," *Int. Labour Rev.*, May 1970, *101*, 475-502.

Nathan Rosenberg, *Perspectives on Technology*, Cambridge 1976.

W. E. G. Salter, *Productivity and Technical Change*, 2d ed., Cambridge 1966.

D. Teece, "Technology Transfer by Multinational Firms: The Resource Cost of Transferring Technological Know-How," *Econ. J.*, June 1977, *87*, 242–61.

DISCUSSION

G. C. HUFBAUER, U.S. Treasury Department: Carlos Diaz-Alejandro sounds a clear lament for the world's poor, but his diagnosis is shadowy. As best I can make out, Diaz-Alejandro believes that the old international economic order (*OIEO*) works to the disadvantage of emerging states. As a consequence, the world trading system—a system that now facilitates the annual exchange of $1 trillion worth of goods—is loaded in favor of economic agents in the northern climes. The rules have worked to nurture monopolies controlled from Europe and North America while fostering competition among the products exported by developing nations. The new international economic order (*NIEO*) would remedy this imbalance through greater "market organization" on the part of developing nations, particularly in primary commodities. So there we have it: "*OIEO* vs. *NIEO*—old monopolists make way for new!"

In answer to these *NIEO* propositions, I would propose a simple experiment for an energetic student:

A. Take a list of commodity groups selected so that there exists a reasonable degree of substitution in production between the commodities within each group.

B. Assume that the market for *any* commodity group can be sufficiently "organized" so that its nominal price can be propelled upwards by as much as a fourfold increase, with no diminution of quantity demanded. (This assumption is unrealistic, but it coincides with *NIEO* thinking, heavily influenced as it is by the success of *OPEC*.)

With only this limit on the reaches of monopoly power, and no requirement of internal consistency in the resulting price structure, the student should try to find a vector of prices that would materially raise the real income of those *states* which house the poorest half of mankind. (This, of course, is an easier test to meet than a test expressed in terms of materially raising the per capita income of individuals making up the poorest half of mankind.)

If such a price vector can be found—a proposition I doubt—then the student should try writing an essay plausibly arguing that the requisite degree of market organization can be achieved for the appropriate commodity groups (without, of course, inspiring the formation of countercartels by producers of the remaining commodity groups).

I submit that this mental experiment will suggest that cartels are not the answer. This does not mean, however, that the poor are forever consigned to misery. Income, after all, is the product of price and quantity. If higher relative prices are not the answer, then larger quantities must provide the source of economic improvement. What is needed is a transformation and diversification of developing economies with greater emphasis on manufacturing.

Consider the following scenario. The *OECD* nations are likely to experience annual real growth of 4 percent annually during the next decade. On past experience, this rate of expansion would support a 7 percent annual growth in world trade; and a 7 percent growth in overall trade would enable exports of manufactures from developing nations to expand at 10 percent annually. Such a rate of growth in manufactured exports could account for 60 percent of the increment in developing nations' foreign exchange earnings during the next decade.

This scenario is by no means assured. Diplomatic efforts directed toward the creation of new cartels in primary commodities may well distract attention from the preservation of a world trading system conducive to the long-term interest of developing nations.

Contrary to the claims of the Group of 77, an expansionist outlook requires not so much a new order as a vigorous defense of the old. The trading system that emerged from the ashes of World War II was ideally, if inadvertently, suited to the export-led prosperity of "two, three, many Japans." This suitability rests on two key features of the General Agreement on Tariffs and Trade: the most favored nation (*MFN*)

principle (Article I, which built on a tradition reaching back to 1417) and the limitation of quantitative import restrictions (Articles XI–XIV).

Successive postwar trade negotiations have witnessed a slow retreat from the *MFN* principle—first the formation of the European Economic Community and other preferential trading groups, then the Long-Term Cotton Textile Arrangement, then the Multi-Fibre Arrangement and the Generalized System of Preferences. Today the Group of 77 clamors for special and differential treatment while European countries press for the right to *selectively* control imports from particular countries whose exports are too competitive.

Such departures from the *MFN* principle make it increasingly difficult for trade to be a vehicle of development. Selective actions often lead to the imposition of quantitative restrictions, which inherently discriminate against the trade of newly competitive and high growth economies. The recent spate of voluntary restraint agreements and orderly marketing agreements is clear evidence of a trend that needs to be reversed.

Both developed and developing countries need to rise up to this challenge if the world trading system is to continue as an "engine of growth" in the 1980's. It is essential that the success countries of the postwar period—Japan, Korea, Taiwan, Brazil, Mexico—open their own markets to imports of manufactures from newly emerging nations. Europe and North America cannot alone bear the costs of domestic economic transformation in response to the trade needs of the developing world.

Developed and developing countries alike must also address the dangers that arise from the proliferation of new techniques of government intervention—tax holidays, outright regional subsidies, and other aids to particular industries. (See Harald Malmgren, *International Order for Public Subsidies*.) Not only can these interventions lead to *inefficient* results; more importantly, they can cause, or appear to cause, the *inequitable* migration of industry—away from areas of comparative advantage. Export-led growth is acceptable, but export-led growth that is built on a regime of public subsidies is not.

In short, the trade challenge is not to form new cartels for primary commodities, but to expand competitive international markets for manufactured goods.

RONALD MCKINNON, Stanford University: In order to better the unsatisfactory lot of the mass of unskilled workers in the third world, Carlos Diaz-Alejandro fervently desires a New International Economic Order (*NIEO*). He would replace (method unspecified) the liberal free-trade principles that have governed most international commerce with and among developed countries in the postwar period.

Anne Krueger has painstakingly assembled data on the labor and skill intensity of export- and import-competing industries for several representative *LDC*s. Paradoxically, her data indicate that past departures from free trade within the Third World may well have substantially *harmed* unskilled workers and contributed to the massive unemployment and underemployment with which we are familiar.

Much of the apparent contradiction arises from Diaz-Alejandro's conviction that effective competition is declining in world markets for goods and services. Producers of industrial products in the advanced countries have increasing monopoly or oligopoly power because they are price setters rather than price takers. This idea should astonish American manufacturers faced with unprecedented competition in textiles, steel, chemicals, machine tools, computers, home appliances, etc., and who have recently been virtually driven out of world export markets for ships, shoes, autos, and other major product lines.

Diaz-Alejandro incorrectly associates increased trade in heterogeneous Hicksian "fixprice" manufactures with increased monopoly power; similarly he associates homogeneous minerals, food crops and natural fibres (Hicksian "flexprice" goods) with the competition that befits the old liberal trading rules. Producers of industrial products with specific brand names quote

stable prices for finite time intervals, measured in weeks or months, so that customers may shop carefully and compare the complex technical specifications of competing products, and incidently hedge the inventories of their distributors against price risk. Texas Instruments quotes a "fixed" but competitive price for its electronic slide rule that is now one sixth of its level three years ago.

A primary product usually sells for a more uniform price that fluctuates on a day-to-day basis but can be hedged in futures markets. However, as Diaz-Alejandro himself points out, this uniformity of price doesn't prevent enterprising monopolists from trying to corner the market for sugar (nineteenth century) or coffee, petroleum, diamonds, and phosphate (this century).

In short, no useful empirical association exists between price rigidity and degree of monopoly power. In today's ferociously competitive industrial environment, those low-wage countries which successfully expand exports (as three or four have already demonstrated) would seem ideally placed to buy high-technology goods at bargain basement prices.

What are the likely internal consequences within *LDC*s of the nascent, but definitely renewed, interest in export-led development? Krueger's study indicates that export industries, which operate mainly at competitive world prices, absorb more labor and less capital—human and nonhuman—than do protected import-substitution industries. If substantiated, this is indeed an important finding.

Heretofore, considerable pessimism has pervaded the extensive empirical studies commissioned by the *OECD* and World Bank regarding the employment consequences of industrialization patterns in the Third World. An unfortunate earlier U.N. initiative, similar to the current *NIEO*, encouraged import substitution in the 1950's and 1960's. Article 18 of the General Agreement on Tariffs and Trade excused *LDC*s from themselves following liberal trading principles. The resulting trade restrictions biased industrial development against unskilled laborers and poor farmers. Fortunately, Krueger's empirical work gives hope that unfortunate income distributions are not endemic to the industrialization process: freer trade policies by the *LDC*s themselves could have more favorable social consequences.

I have only one specific criticism of Krueger's interpretation of her data. She presumes the higher capital-labor ratios of inter-*LDC* exports vindicates the factor proportions theory of comparative advantage. More likely, this is the artificial result of customs unions (for example, the Andean Pact) or bilateral trade agreements (say in automobile components) among *LDC*s that encourage the bartering of protected products that are not meeting the market test of world prices. Perhaps her export data could be further broken down to suppress this effect.

In summary, Krueger and associates have tentatively shown that if *LDC*s dismantle their trade restrictions, then the resulting industrialization patterns might yield tendencies toward factor price equalization which would improve the lot of unskilled labor in the Third World, much more than seemed possible based simply on the industrialization experience of the 1950's and 1960's. For this to occur, developed countries must, however, strongly resist protectionist pressure for government "management" of their imports of goods and services from *LDC*s—something the *NIEO* enthusiasts seem perversely close to advocating. And of course Diaz-Alejandro is correct that the more liberal movement of people, through reduced restrictions on the migration of unskilled labor, is an important adjunct to free commodity trade.

ROBERT E. BALDWIN, University of Wisconsin-Madison: As Carlos Diaz-Alejandro points out in his very interesting paper, the basic complaint of the developing countries is that they feel that the markets in which international transactions take place are somehow rigged against them. What this invariably comes down to is that they believe the developed countries use their greater economic, political, and even military

power to extract sizable monopoly rents from them. There is no doubt, as Diaz-Alejandro also points out, that international markets for both minerals and many foodstuffs were far from purely competitive even in the last century. However, I do not think too many economists have ever thought they were. It is perhaps a bit like setting up a strawman to imply that economists in developed countries thought that atomistic competition was the general practice, whereas in fact it is very easy to cite many cases where clearly this was not true.

Be that as it may, I think the point Diaz-Alejandro is making is absolutely right. Certainly until after World War II, when many *LDC*s gained their political independence, the monopoly power that existed in international economic relations was mainly in the hands of economic agents of the developed countries and they did not hesitate to use it. However, from the 1950's onward, this situation began to change considerably. The newly formed nations among the *LDC*s as well as some of the older independent developing countries began to use the power of the nation-state to extract some of the monopoly rents from foreign investors in their countries. In the mineral field, for example, there was first a sharp rise in required royalty and tax payments and then more and more nationalization proceedings against foreign companies. As of now, I doubt if there are many more rents that can be extracted in the mineral area, and we know, of course, that mineral exploration has shifted sharply away from the developing countries towards developed countries because of some of these actions.

What we see happening now and what is being pressed for under the *NIEO* carries these actions a step further. Competition among producers of commodities like copper or oil did not produce the competitive solution but neither did it produce the monopoly one. Efforts of the kind of strict collusion required for the monopoly solution invariably broke down. However, as these oligopolistic private producers are being replaced by even a smaller number of governmentally controlled producers, opportunities exist for the developing countries to extract a greater rent by coming closer to the pure monopoly solution. The realization of this power and its use, as in the oil case, can produce profound shifts in income distribution. Where the developing countries see that the OPEC's solution will not work, they are seeking other institutional means such as the use of buffer stocks and price fixing schemes designed to stabilize and raise prices of various primary products.

These efforts really do not represent claims for a new order but rather efforts to use the old order for monopolistic purposes just as agents from the developed countries did and do today. The *LDC*s can now do this because by acting on a governmental level they are large enough to exercise real monopoly power. We should not be surprised by this nor claim they are now rigging the rules against the developed countries any more than we should say the developed countries formerly rigged the rules against the developing countries. In an environment in which antimonopoly rules are not administered on an international basis, both of these cases are merely developments under the existing rules.

Whether we can say—as Diaz-Alejandro at least hints—that the new monopoly power of the *LDC*s will act as a counteracting force against the *DC*s that actually improves the allocation of resources is not clear. From my viewpoint it would seem that when new monopolies arise to face old ones, the most frequent result is an even poorer allocation of world resources. Of course, ideally what we need are a set of international rules that discourage monopoly, whether it be on the part of *DC*s or *LDC*s, coupled with transfer mechanisms that handle our concerns for equity. But just as these rules were not politically feasible when the *DC*s exercised most of the monopoly power, it is unlikely that they are politically feasible now. So we shall be left with a situation in which the *LDC*s are increasingly able to meet their equity goals by exercising monopoly power—unfortunately at the cost of greater inefficiency of world resource allocation.

Consumer Product Safety Regulation

By HENRY G. GRABOWSKI AND JOHN M. VERNON*

Government policy toward consumer product safety has experienced major institutional changes in the United States over the past fifteen years. In particular, Congress has passed a number of laws imposing and strengthening federal regulatory controls on product safety across a broad spectrum of markets.

In the food and drug area, the 1962 Kefauver-Harris Amendments made the premarket approval process for new pharmaceuticals much more stringent and extended Food and Drug Administration (*FDA*) controls over the pharmaceutical research and development process. The 1968 Delaney Amendments required the *FDA* to ban any food additive from the marketplace found to be carcinogenic in animals, regardless of foregone benefits. The 1976 Medical Device Amendments extended *FDA* controls to all medical devices (for example, heart pacemakers, cardiographs, stethoscopes, etc.) and many classes of medical devices will now be subject to a premarket approval process similar to that for new drugs.

Beginning in the mid-1960's, Congress also has passed a succession of product safety laws dealing with specific products such as automobiles, toys, flammable fabrics, lead-based paints, and poisonous and toxic substances. Most of these responsibilities were eventually consolidated and put under the jurisdiction of the Consumer Product Safety Commission (*CPSC*), created in 1972. This new agency was given a broad mandate by Congress to set safety standards for all consumer products presenting undue risk of injury, except for those products already regulated by an established agency (for example, food, drugs, pesticides, and autos).

*Duke University.

Thus Congress has extended federal product safety controls to virtually all areas of the marketplace. While it is too early to evaluate the impacts of this new regulation, it is possible to make some general observations about its emerging characteristics.

First, in drafting and funding new product safety legislation, Congress has strongly favored direct regulatory controls (for example, product standards, premarket approval, prohibitions of very risky products, etc.) compared to other policy instruments that might be employed to encourage greater product safety. In particular, two alternatives often advocated in the academic literature—the generation and dissemination of better information about product safety hazards and the use of economic incentives (i.e., taxes or subsidies) have been given little attention.

Second, the decision-making process at the various agencies appears to embody a strong "safety imperative." That is, there is strong resistance to the notion that the benefits of greater safety stemming from a particular policy must be weighed against the costs that might be entailed by that policy. To a considerable degree, the regulatory agencies are probably reflecting the desires of Congress in this regard. The product safety laws tend to be drawn with very specific and narrow mandates (for example, to protect consumers against unsafe products) and provide few incentives for agency decision makers to introduce cost considerations into their decisions. While it is true that these agencies are now required to calculate "economic impact" or benefit-cost analyses of their decisions, these generally take on an "after the fact" character. As we show in our analyses of the *CPSC* and the *FDA* below, the results of benefit-cost analyses apparently have little effect on regulatory decisions.

Third, there currently exists little effort to design regulatory policies so as to complement existing market and legal incentives regarding product safety. Presumably the rationale for government regulation of product safety rests first on the presence of market information imperfections; and second, on the fact that the tort liability system (which makes producers liable for "defects" in their products) provides weak incentives in many circumstances because of high transactions costs and uncertainties in legally determining fault. However, these market and legal imperfections vary greatly across different product areas and industry categories. Therefore, in setting priorities for regulatory action, an agency with broad discretionary powers like the *CPSC* should presumably concentrate on those areas where market and legal incentives for product safety are most deficient. In this way they can target their resources to the areas where potential benefits are greatest relative to costs.

Fourth, product safety standards and regulations can result in significant unintended side effects on the long-term competitive structure of an industry. Our own recent analysis of the pharmaceutical industry, for example, indicates that increased regulation since 1962 has resulted in a much greater concentration of innovation among the largest drug firms. Similarly, recent analyses of the *CPSC* proposed standards in power lawnmowers indicate that implementation of these standards would eliminate several small producers and significantly increase industry concentration (see W. Brockett et al.).

In the remainder of the paper, we specifically analyze the behavior and performance of the two principal agencies engaged in product safety regulation (the *CPSC* and the *FDA*) and consider these points in more detail.

I. The Consumer Product Safety Commission

Congress passed the Consumer Product Safety Act in 1972 which created the *CPSC* and empowered it to "protect the public against unreasonable risks of injury associated with consumer products" (*CPSC*, 1976). It has been estimated that the Commission has jurisdiction over some 10 to 11 thousand different products which account for about $750 billion in annual sales. Among the policy options which the Commission has to carry out its mandate are the dissemination of information, the development of minimum safety standards, and the outright ban of especially hazardous products.

To date the Commission has proposed or implemented safety standards for products such as bicycles, matchbooks, power lawnmowers, swimming pool slides, and public playground equipment. In establishing priorities on standards, the Commission has relied heavily on its frequency severity index of product related injuries. This is based on the number and character of injuries from a particular product class recorded at hospital emergency rooms.

This approach to establishing agency priorities has been strongly criticized in a recent study by Nina Cornell, Roger Noll, and Barry Weingast. In particular, they argue that while the products targeted for standards by the *CPSC* have above average injury rates, they are products involving risks which are well understood and voluntarily assumed by consumers. At the same time, the Commission has given little attention to more sophisticated products like microwave ovens, where the hazards are more subtle and less clearly defined and for which information on safety characteristics is more difficult for consumers to obtain. This type of product, of course, rarely shows up as the cause of emergency room injuries, but may pose significant long-run health hazards about which there is general consumer ignorance.

In our opinion, a major ongoing problem with the *CPSC* approach to product safety regulation is that it does not really try to compare benefits and costs in deciding where government safety standards are necessary. Rather, the Commission's decisions reflect a "safety imperative" which tends to ignore the cost side of the equation

almost completely. This is demonstrated by an analysis of the priority rankings for forty-six product classes which are considered in the *CPSC Mid-Year Review* (March 1977). This report suggests various factors and criteria as relevant to establishing Commission priorities including the frequency and severity of injuries, causes of injuries, costs and benefits of *CPSC* action, unforeseen nature and vulnerability of the population at risk, the probability of exposure to hazard, and other factors.

Table 1 presents the rankings and corresponding benefit-cost ratios for twenty-one product classes for which ratios were available. Although many of these benefit-cost ratios are based on very preliminary economic analyses, they are the numbers available to the staff and commissioners in establishing priorities. The priority rankings refer only to the Commissioners' ordering of the twenty-one products in Table 1, and not their rankings among all forty-six products that they evaluated. However, rankings 1–12 in the table were all accorded the status of "high priority" and are targeted for standards by the *CPSC* during the coming year.

TABLE 1—*CPSC* PRIORITY RANKINGS AND BENEFIT-COST RATIOS OF CURRENT AND FUTURE PROJECTS

Project	Benefit-Cost Ratio	Priority Ranking
Bathtubs and Showers	2.70	12
Over-the-Counter Antihistamines	2.52	15
Public Playground Equipment	2.02	3
Gas Space Heaters	1.85	2
Drain Cleaners	1.08	17
Ladders	.94	11
Glazing Materials	.91	4
Ranges and Ovens	.87	7
Trouble Lights	.75	8
Chain Saws	.67	14
Upholstered Furniture	.48	5
Power Saws (portable)	.40	13
Power Mowers	.40	1
Matches	.37	10
Rust Remover	.34	20
Petroleum Distillates	.25	16
Power Saws (nonportable)	.16	18
Ammonia	.11	21
Extension Cords	.10	8
Television Sets	.09	6
Wearing Apparel	.02	19

Source: Benefit-cost ratios were obtained directly from a *CPSC* Bureau of Economic Analysis staff memorandum or calculated from new or revised data supplied in the *CPSC Mid-Year Review* using the procedures of the *CPSC* Bureau of Economic Analysis; priority rankings were obtained from a *CPSC* News Release.

It is clear from Table 1 that the *CPSC* does not accord great weight to benefit-cost analysis, either in the absolute sense of the desirability of pursuing the project at all or in the ranking of projects. Only five projects of the twenty-one have ratios exceeding unity. Furthermore, the number one priority ranking in the table, power mowers, was ranked second by the Commissioners out of forty-six and it has a benefit-cost ratio of only .40 (total benefits were estimated at $112 million compared to costs of $285 million).

In doctoral dissertation research currently underway, Lacy Thomas is analyzing the *CPSC* decision-making process. His effort is directed at determining empirically the implicit weights for project attributes that the *CPSC* uses in establishing its choices among projects. Using a logit analysis, Thomas has found that estimated benefits (which are highly correlated with the estimated frequency and severity of injuries for each product class) dominate cost considerations in the setting of agency priorities. In particular, estimated coefficients on the benefit variable are ten to twenty times larger in absolute magnitude than those on the cost variable.

It should be noted that *CPSC* members and other product safety regulators have argued that there are very good reasons for not making their decisions depend directly on the outcomes of benefit-cost analysis. First, they suggest there is no generally accepted operational methodology among economists for valuing human lives. Second, they point out that the benefits and costs are not comparable. The benefits involve the saving of human lives and the reduction of bodily injuries and health hazards while the costs involve higher

product prices, lower business profits and other "economic" effects.

While these arguments are apparently quite persuasive to many congressmen and consumer advocates, at best they only argue against the use of a strict benefit-cost criteria of unity in accepting or rejecting a project. It it still appropriate to use some type of benefit-cost calculation in ranking projects and setting priorities, if the Commission's actions are to be *cost effective*. The rankings in Table 1 obviously do not have this property. In effect, they would seem to imply that the benefits of saving lives or preventing injuries for a product class like television receivers or extension cords (which are given high priority but have very low benefit-cost ratios) are worth several times more than the corresponding benefits obtainable from product classes like chain saws or drain cleaners (which have much higher benefit-cost ratios but are given considerably lower priority by the Commission).

The utilization of benefit-cost analyses in this manner would also seem to have advantages in helping the Commission to choose among alternative strategies of government intervention. There may well be cases, for example, for which safety standards have a relatively higher benefit-cost ratio but where some alternative strategy (for example, information dissemination) could accomplish the same objective in a more cost efficient manner. While the Commission and Congress have tended to favor safety standards over other strategies, at present they cannot justify these preferences on cost efficiency grounds since the requisite benefit-cost analyses have never been undertaken.

II. FDA Regulation of Pharmaceuticals and Medical Devices

In contrast to the recent charter of the *CPSC*, government regulation of pharmaceuticals started in 1906 and has evolved over time into a very stringent system of premarket controls over new drug development and introduction. While early regulation was oriented at patent medicine abuses, the sulfanilamide tragedy in 1938 led to passage of the Food Drug and Cosmetic Act which required *FDA* approval of all new drugs as "safe" before they could be marketed. Then in 1962, as the disastrous effects of thalidomide were becoming apparent in Europe, the Kefauver-Harris Amendments were passed. This law expanded *FDA* controls to the clinical testing and development process for new drug compounds. In addition, manufacturers were required to demonstrate the therapeutic efficacy as well as safety of a new drug to obtain *FDA* approval.

The fact that new drugs can be the source of serious unforeseen toxic side effects in addition to strong therapeutic benefits justifies these strong regulatory controls in the minds of many individuals. At the same time, *FDA* regulatory decisions have been characterized by an extreme form of safety imperative. As *FDA* Bureau of Drugs director Richard Crout has indicated: "I would emphasize very strongly that the Food and Drug Administration regulates health policy, not economic matters. That is terribly important to understand. We do not pay any attention to the economic consequences of our decisions and the law does not ask us to" (pp. 196–97).

Over the period since the 1962 Amendments were passed, a number of adverse trends have been observed with regard to the innovative performance of the pharmaceutical industry. In particular, average research and development costs for a new drug entity have increased more than an order of magnitude and now exceed $20 million per new drug. Development times and risks have also significantly increased. Most importantly, the annual rate of new drug introductions in the United States has fallen to less than one-third the rate which existed in the early 1960's. In a forthcoming paper we consider various hypotheses for these adverse trends; and, on the basis of international comparative analysis, conclude that increased regulation has been a major factor underlying declining innovational performance in the drug industry (see

the authors and Thomas). This is consistent with a number of other studies (see Martin Baily, Sam Peltzman, David Schwartzman).

There are clearly foregone health benefits to the public when beneficial drugs are left undeveloped or are substantially delayed because of *FDA* regulatory controls. William Wardell, a clinical pharmacologist, and Louis Lasanga documented many cases in which new drugs developed abroad (and even many American drugs first introduced abroad) generally took several additional years to gain *FDA* approval for use in the United States. Their findings are consistent with our own analysis of the international diffusion of new drug therapies across four countries (the United States, United Kingdom, France, and Germany). Specifically, we found that a majority of all the new chemical entity drug introductions into the United States over the period 1965–75 had a prior introduction in the United Kingdom, France, or Germany. Moreover, if one considers only the twenty-seven new drugs introduced in this period that were specifically classified by the *FDA* in 1974 as *important therapeutic advances*, fifteen had prior introduction in one of these foreign countries, eight became available here and abroad in the same year, and only four were initially available here first.

Of course, the total benefits of *FDA* regulation may exceed its overall costs. We have been concentrating here on the cost side, and the benefits to the public of avoiding harmful side effects of drugs kept off the market surely exist. However, given the *FDA*'s stated policy of giving absolute priority to considerations of safety and efficacy, and ignoring effects on firm costs and innovation, it is almost inevitable that *marginal* costs will exceed benefits and that *FDA* policy will err on the side of being overly restrictive.

As noted in the introduction, Congress, in the Medical Device Amendments of 1976, has extended the *FDA* premarket regulatory controls over a larger spectrum of medical products. If the *FDA* brings a similar regulatory philosophy to bear on this sector to that which it has exhibited in pharmaceuticals, the costs in terms of foregone innovation are likely to be quite high indeed. This is particularly so because innovation in many medical device fields (such as heart pacemakers) has not been characterized by large economies to scale and several major new products have emanated from small firms. Such firms would be least able to finance or bear the costs and risks of an expensive, lengthy, and uncertain premarket regulatory approval process. Moreover, we have shown elsewhere (1976) the rapid increases in research and development costs that occurred in pharmaceuticals over the post-Amendment period has operated to concentrate innovation in the very largest drug firms. One might expect similar, but perhaps even more dramatic structural changes for many medical devices, if regulation in this area proceeds with a comparable approach to *FDA* regulation of pharmaceuticals.

III. Summary and Conclusion

While it is still too early to evaluate conclusively the recent wave of consumer product safety regulation, the evidence thus far indicates that serious resource misallocation is taking place and is likely to continue. The regulators of product safety tend to rely solely on direct controls (product bans and standards) and to be concerned with the benefits only, as measured in the number of lives saved and accidents avoided. They ignore, often intentionally, the costs of their controls. Unless this "safety imperative" approach to regulation is changed, the problems of resource misallocation will multiply over time as regulatory controls are extended to several additional product classes.

REFERENCES

M. N. Baily, "Research and Development Costs and Returns: The U.S. Pharmaceutical Industry," *J. Polit. Econ.*, Jan. 1972, *80*, 70–85.

W. Brockett et al., *An Analysis of the Pro-*

posed CPSC Lawnmower Safety Standards, Stanford 1977.

N. W. Cornell, R. G. Noll, and B. Weingast, "Safety Regulation," in Henry Owen and Charles Schultze, eds., *Setting National Priorities: The Next Ten Years*, Washington 1976, 457–504.

J. R. Crout, "Discussion," in Robert B. Helms, ed., *Drug Development and Marketing*, Washington 1975, 196–97.

H. G. Grabowski and J. M. Vernon, "Structural Effects of Regulation on Innovation in the Ethical Drug Industry," in Robert T. Masson and P. David Qualls, eds., *Essays on Industrial Organization in Honor of Joe S. Bain*, Cambridge 1976, 181–206.

———, ———, **and L. G. Thomas,** "Estimating the Effects of Regulation on Innovation: An International Comparative Analysis of the Pharmaceutical Industry," *J. Law Econ.*, forthcoming 1978.

S. Peltzman, "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments," *J. Polit. Econ.*, Sept. 1973, *81*, 1049–91.

David Schwartzman, *The Expected Return from Pharmaceutical Research*, Washington 1975.

L. G. Thomas, "A Statistical Analysis of the Priorities of the Consumer Product Safety Commission," mimeo., Duke Univ. 1977.

William M. Wardell and Louis Lasagna, *Regulation and Drug Development*, Washington 1975.

Consumer Product Safety Commission, "BEA Summary Table," Bureau of Economic Analysis staff memo., mimeo., July 1976.

———, *Mid-Year Review*, Washington Mar. 1977.

———, *News Release*, no. 77-55, June 3, 1977.

Egonomics, or the Art of Self-Management

By T. C. SCHELLING*

One of the sophisticated financial arrangements available at your neighborhood bank is "Christmas Savings." In this plan you are committed to regular weekly deposits until some date in November when all the money is there with accumulated interest to spend for Christmas. It doesn't accumulate quite as much interest as regular savings. The reason people accept less interest on Christmas savings is that the bank protects these funds a little more than it protects ordinary savings. Regular savings are reasonably well protected against robbery, embezzlement and insolvency; and insurance takes care of what protection cannot do. But there is one predator against whose ravages the bank is usually impotent—you. With a Christmas account, the bank assumes an obligation to create ceremonial and administrative barriers to protect your account from yourself.

Some people cheat on the withholding-tax forms they fill out for their employers. They understate their dependents, so that the Internal Revenue Service takes more than it deserves all year—a free loan from the taxpayer—in return for which the taxpayer gets a reduced shock the following April.

Many of us have little tricks we play on ourselves to make us do the things we ought to do or to keep us from the things we ought to foreswear. Sometimes we put things out of reach for the moment of temptation, sometimes we promise ourselves small rewards, and sometimes we surrender authority to a trustworthy friend who will police our calories or our cigarettes. We place the alarm clock across the room so we cannot turn it off without getting out of bed. People who are chronically late set their watches a few minutes ahead to deceive themselves. I have heard of a corporate dining room in which lunch orders are placed by telephone at 9:30 or 10:00 in the morning; no food or liquor is then served to anyone except what was ordered at that time, not long after breakfast, when food was least tempting and resolve was at its highest. A grimmer example of a decision that can't be rescinded is the people who have had their jaws wired shut. Less drastically, some smokers carry no cigarettes of their own, so they pay the "higher" price of bumming free cigarettes.

In these examples, everybody behaves like two people, one who wants clean lungs and long life and another who adores tobacco, or one who wants a lean body and another who wants dessert. The two are in a continual contest for control; the "straight" one often in command most of the time, but the wayward one needing only to get occasional control to spoil the other's best laid plan.

As a boy I saw a movie about Admiral Byrd's Antarctic expedition and was impressed that as a boy he had gone outdoors in shirtsleeves to toughen himself against the cold. I resolved to go to bed at night with one blanket too few. That decision to go to bed minus one blanket was made by a warm boy; another boy awoke cold in the night, too cold to retrieve the blanket, cursing the boy who had removed the blanket and resolving to restore it tomorrow. The next bedtime it was the warm boy again, dreaming of Antarctica, who got to make the decision, and he always did it again.

I didn't realize then how many contests of that kind, some pretty serious, I would eventually have with myself, trying to stop smoking, to exercise, to study for an examination, to meet a deadline, or to turn off an old movie on TV. At a gathering like the annual meeting of the American Economic Association most of us are exquisitely aware of that form of academic delinquency that is probably our greatest occupational hazard: We cannot make ourselves write those papers, articles, and dissertations

*Harvard University.

that we know we must write, the impediment being greater the more important the thing we have to write. In the end it is often deadlines, sometimes deliberately contrived for the purpose, that enhance our desperation to the point where the risk of failure can no longer keep us from trying. Some of my colleagues have told me of the artful ways they contrive to make themselves get started on that paper, or how they break a frightening large task into small pieces with a rule that one piece must be done each day.

Some of our contrivances are ingenious and successful. If told by a doctor we'd live longer if we'd get out in a cement covered yard and jump up and down for an hour, most of us would settle for shorter lives; but if we get a ball and something to hit it with, and somebody to hit it back, and make rules to convert the jumping into a contest, the activity becomes quite engaging. (Indeed, some even become excessively engaged.) I run for exercise and, I believe like most people, I dislike it; I keep waiting for the inventor of that mechanical rabbit at the dog races to contrive something that adds comparable zest on a people's course.

Many sophisticated people use unsophisticated budgeting devices, like weekly spending allowances and special funds for self-indulgent expenditures, often using two or three savings accounts, sometimes in separate banks, to achieve self-control through self-intimidation. They cannot casually erode a boundary they have put around certain expenditures if they have to violate a categorical rule or walk into a separate bank to steal their own money. Recently Henry Rowen and Jess Marcum have offered an interpretation of betting in lottery-like games with very long odds. "For people who play these games, savings would be a more reliable way of amassing some money," because of the actuarial unfairness of the odds; but if these are people who cannot discipline themselves to accumulate large sums through a protracted period of savings, and will never have a lump sum with which to escape from their low-capital life style, "such games might seem to offer the only hope of accumulating a lump sum." And a good part of the rationale for social security and mandatory retirement plans has been that people will be better off if they are obliged to do what they would usually wish to do but for which they would suffer occasional disastrous lapses of morale and self discipline. Sidney Alexander once told me, after leading a research project for several years, that he had finally learned what an entrepreneur is: a person who spends most of his time getting people to do what they said they would do. And the rest of us spend a good part of our time trying to get ourselves to do what we already decided to do.

Surveys in America and in England indicate that most people who smoke—by no means everybody, but a majority—have at some time tried to stop. The Surgeon General has been warning people for two decades that smoking is bad. Just about everybody knows it. If there were some way that cigarettes could be reliably put beyond reach, and people could vote on whether they would like that done, it is a fair guess that a majority of the smokers would elect to deny themselves any possibility of lighting another cigarette.

Hardly anybody thinks it could be done, and neither experience with alcohol in the 1920's nor marijuana in the 1960's makes the effort look promising. Those who didn't want the cigarette ban would offer a market for contraband cigarettes; nobody has a good idea how to suppress such a market; and, once the market is there, the smokers who favored the ban will be little more able to resist cigarettes than they used to be. And even if the abolition were unanimously approved by all smokers, people would know that if they could sneak in a few cigarettes people would buy them; there would be a black market and most of the people who wished the market didn't exist would patronize it.

Smoking is only one of several addictive or habitual behaviors that people engage in, but it is the best example of one that is widespread, meets no known physical need

(except for people who have already acquired the habit), is known to be harmful but only in the statistical long run, is hard to quit, and one that most people might like to quit, especially if they could be relieved of withdrawal difficulties but even if they had to suffer if only they were assured of success. Overeating is second in the number of people who wish they could control their behavior better than they do; and alcohol has a large absolute number, whether or not it comes close to being a majority who would quit altogether if only they could, as the only way of bringing their consumption under control. There are also people who gamble and wish they didn't, or watch too much television. The phenomenon of addiction is widespread, and by no means should all addictions be deplored, nor are all those we deplore necessarily candidates for any kind of action. (If people are addicted to exercise it may be great for them, though it infuriates their friends.)

It is not clear with marijuana, as opposed to tobacco, that a large part of the participants are convinced usage is bad for them, wish they could quit, and can't. I am distinguishing the drugs on one side and the food, tobacco, and possibly alcohol on the other, because they represent altogether distinct issues in "social control." Nearly everybody who wants heroin suppressed is not an addict. And not many people who take heroin are pleading to be deprived of it. But tens of millions of people wish they could smoke less, or quite smoking, and the primary constituency for social action against cigarettes is probably not among the nonsmokers, but among those who smoke. (This undoubtedly includes some who smoke but are more concerned about their children than about themselves.)

There is a recent innovation that we can watch with interest. In some cities you can now go to a shopping mall and subscribe to a commercial program to help you stop smoking. The idea makes sense: most of us have little knowledge of the ways that people bring a habit or addiction under control; help would be worth money. (Even the savings on cigarettes alone are more than enough to pay for the course.)

Actually, all is not hopeless. Since the Surgeon General's findings were first made public in the 1960's, the number of cigarettes per capita stopped increasing and has decreased slightly. The tar content has declined markedly. In my own census group, males 45 to 65, the proportion that smokes is declining by 4 percent a year. In this age group, white male professionals have most strikingly discontinued smoking. You can verify that by looking around you at the annual meeting of the American Economic Association: cigarette butts per capita in the ashtrays is a small fraction of what it was ten years ago, and if you check the brands you will find tar and nicotine down another half.

Little is known about how these people stopped smoking—and stop they did, at least many of them, because there were not only fewer smokers but more "former smokers" among those middle aged males. (The figures can be confusing: among women the smokers are still increasing while former smokers are increasing almost as rapidly; and among middle aged women smokers are increasing slightly and former smokers are increasing much more rapidly.) We have a little information on how many cigarettes per day the smokers smoke, but no information on how many times the "former smokers" have quit and started again, or even how many of them just hadn't had a cigarette for a whole day when the interviewer rang the bell.

What is it worth to quit smoking? Suppose there were a reliable way to quit, and to quit wanting to smoke, without torment or suspense or loss of dignity or any physical side effects. What would it be worth to those 50 million smokers, and to those 30 million former smokers of whom some are going to need it and a few could use it right now? Let me conjecture an immediate market of 30 million customers. If those who would like to quit, smoke just as much as those who are not interested, these 30 million smokers spend about $10 billion per year on cigarettes. (A third of this is a transfer, by the tax system, to nonsmokers and to each other, so our remedy for smoking is also a tax relief device.) If smokers

expect, in the absence of relief, to smoke another fifteen years or more, and if they discount future savings at somewhere from 8 to 12 percent per year, and if at a minimum they would value relief from smoking the way they would value the fuel oil savings from warmer weather, we can put a minimum valuation somewhere from $75 to $100 billion. (There will be some "producer's surplus" in this figure—farmers and retailers to whom reduced sales would represent a net loss, so the net social saving after taxes will be somewhat more than half what the consumers save.)

We can only guess what people would pay to be relieved of the nonfinancial cost associated with smoking—better health, freedom from a "habit," cleaner teeth, or cleaner ashtrays—and what people would pay to help spouses, children, parents, and friends to be rid of what is usually considered an unhealthful addiction. We don't know because they don't know. (Some of "us" are among "them" and we still don't know.) Many of them, furthermore, prefer not to face the question, especially as the guaranteed torment-free treatment is not now for sale. Moreover some would probably insist on stopping (or failing) on their own, and not by the procurement of artificial support for their self control.

An alternative question is how much smokers would pay for something that, with little impairment of their smoking pleasure, would make the habit perfectly safe, and certified so by that Surgeon General who otherwise tells us that smoking is dangerous.

My conjecture, which you may compare with your own, is that the worth to consumers of being free of smoking, or free of the consequences of smoking, is greater that the gross financial savings. Then the total "consumer's surplus" from suspense-free torment-free nonconsumption of cigarettes, discounted to the present for today's smoking population, is in the neighborhood of a quarter trillion dollars.

I want to draw the conclusion that this is an anomaly in consumer theory, consumers getting negative satisfaction out of something they spend a lot of money to consume. But life is full of expensive things that we want, the relief of wanting which would save us lots of money; and wishing that we didn't want to smoke may not be different from wishing we didn't get cold or wishing we didn't need sunburn lotion or novocaine. Even the argument that smoking is an addictive habit that people acquire thoughtlessly or irresponsibly, and that they would be better off had they been denied the opportunity to become addicted, is not altogether different from the argument that people have bad teeth and gums because they were thoughtless and irresponsible about dental hygiene earlier in life and might be better off had they been compelled to incur the nuisance of oral hygiene.

To establish the anomaly, I must conjecture a stronger proposition, one that I believe could be demonstrated but has not been: among those interested customers seeking a way to stop smoking, most of them would settle for a technique that guaranteed that cigarettes would not be available when they wanted to smoke. This technique would be suspense free but not torment free. All the withdrawal symptoms except those arising in doubt, indecision, suspense, and guilt would be accepted. Something that would make cigarette smoking instantly painful—as in an alternative system children's teeth might become instantly painful upon failure to brush them properly—would be accepted by people for themselves, according to my proposition, as the other might be accepted by parents for their children's teeth.

I'll go further and offer an expectation—not part of the proposition that I ask you to entertain—that the reluctance of some of our potential customers to face the pains of cigarette starvation would diminish rapidly as they discovered, through the experience of others, that even the withdrawal symptoms were not all that bad unless aggravated and prolonged by doubt, indecision, or occasional relapses. (There is some observation that people who discover that they absolutely *must* quit smoking and who know that therefore they *will* quit smoking

find the withdrawal difficulties of less intensity and shorter duration, sometimes dramatically less, than even their own experience had led them to expect.)

Meanwhile, with no visible encouragement from the Surgeon General, cigarette companies in their advertising are engaged in a spectacular competition for the low-tar market, a market that may actually be expanding explosively only because, still without quite admitting that tar may cause cancer, cigarette companies are beginning to treat "tar" as though it is a pollutant. Compared with the puny commercial efforts to merchandise courses on how to stop smoking at the local shopping mall, the several hundred million dollars a year spent competitively advertising cigarettes in newspapers, magazines and billboards may end up doing more to reduce the harmful effects of smoking than consumers have been able to do for themselves.

Safety Decisions and Insurance

By MARTIN J. BAILEY*

Just as choices among gambles and among insurance policies can in principle reveal a cardinal utility function unique up to a linear transformation, as was pointed out by Milton Friedman and Leonard Savage, so also can choices between safety and income. For persons with no assets who buy insurance we can show a priori that the "value of life" implicit in safety choices exceeds discounted lifetime earnings. The lower bound for this value is a simple function of the loading charge of the insurance and the amount of insurance purchased.

I. Revealed Preference for Safety vs. Income or Consumption

Several authors have explored the implications for the value of life of choices between safe and hazardous jobs, work which I recently summarized (1978). Glenn Blomquist analyzed the use of automobile seat belts to infer such a value. In addition to these, there are in principle many other choices people make that contain information about their willingness to pay for greater safety: purchase of safety devices such as smoke detectors; vaccinations; medical checkups; other services of preventive medicine; and so on.

For choices that exchange income (or consumption) for safety, it is now an established convention to denote the derivative of income with respect to personal risk as the value of life—the willingness to pay, or compensating variation, for small changes in safety. One can visualize this concept by considering a group of 1,000 workers, each of whom would be marginally willing to accept a job with an extra risk of one death per 1,000 workers per year, with extra net after-tax wages of $200 per worker per year; however, they happen to have taken less risky jobs at lower pay instead. As a group, these workers have given up total wage income of $200,000 per year, and they suffer one less death per year (say, 6 deaths instead of 7 in the group of 1,000) in their chosen jobs. Thus they are willing to pay $200,000 to have one less death in their midst, or $200,000 "per life saved."

Some of the choices involve changing measured income in exchange for changes in safety; others involve spending out of income, that is, reducing consumption but not income for this purpose. The treatment of this difference has not had the benefit of a clear convention; for example, some authors show utility as a function of income, others as a function of consumption. It will be helpful to clarify this point.

Although a lifetime has many stages or periods and many risks, we lose no generality for present purposes if we use a choice model with a single period in which the household uses lifetime income to buy lifetime consumption (including bequests, if any). The household maximizes the Lagrangian

$$(1) \qquad U^* = U(C) + \lambda[Y - C]$$

Because at the maximum of U^* for each Y the household sets $Y = C$, we can write

$$(2) \quad U^{**}(Y) = \max U^*(C, Y) = U(Y)$$

so that for most purposes, C and Y are interchangeable as the argument of U^{**} or U.

However, there are significant exceptions to this interchangeability. For example, consider a Friedman-Savage gambler whose utility function is convex in the large for the range of values of C from C_0 to C_1, as shown in Figure 1. If he can lend, borrow, or gamble, this person can spend part of his life consuming C_0 and part consuming C_1; in the case of gambling at fair odds his expected utility is a probability weighted

*University of Maryland.

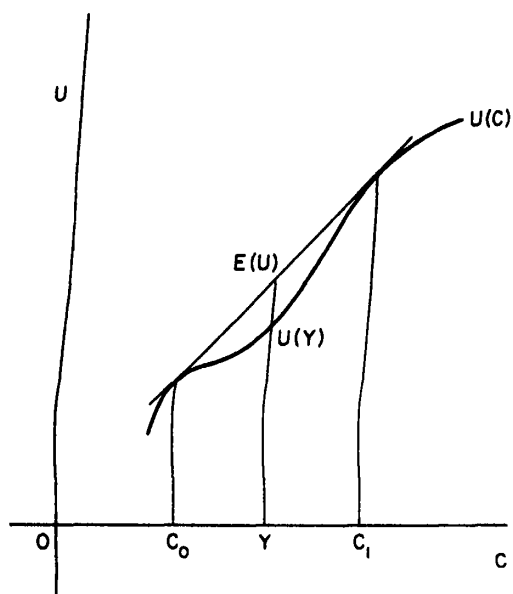



FIGURE 1

average of $U(C_0)$ and $U(C_1)$, which is linear in Y, with

(3) $$\frac{dE(U)}{dY} \equiv U'(C_0) \quad C_0 < Y < C_1$$

Hence, the "utility of income" in this case is the linear function $E(U)$ over the range of income in question, and is greater than the utility of consumption in this range. (The same result holds if the person can borrow and lend.) For this case the utility of income or the expected utility of consumption is the pertinent concept for the willingness to pay for safety. When utility is concave, however, consumption is unambiguously the argument of the utility function, although income may take its place if consumption equals income.

II. The Value of Life and the Availability of Insurance

I now turn to the main issue: the effect of insurance on the willingness to pay for safety. The preceding section establishes that we can assume that $E(U)$ is semiconcave; for simplicity, let $U(\ )$ be strictly concave. Let us assume throughout that a casualty event (bad news) reduces lifetime

income to zero, unless the victim has a insurance policy, and may also take hi life. Buying added safety protects agains both losses, whereas insurance protects only the income. We also assume that the purchaser of insurance or of safety pays only i there is no casualty (good news). For the uninsured person this assumption is material; it is proper here because we focus attention on casualties that involve total loss of income. For this assumption, if good news has the probability p and bad news the probability $1 - p$, expected utility for the uninsured person is

(4) $$E(U) = pU(Y) + (1 - p)\,U(0) = pU(Y)$$

If the casualty can be fully insured—it doesn't kill the victim if he doesn't starve—he can buy insurance with bad news payoff of C' for a premium Q, to obtain the expected utility

(5) $$E(U) = pU(Y - Q) + (1 - p)\,U(C')$$

Let the insurance be fair, in which case

(6) $$Q = \frac{1 - p}{p} C'$$

Maximizing $E(U)$ leads to full coverage of $Y - Q$ (see Isaac Ehrlich and Gary Becker), so that with (11) we have

(7) $$Y - Q = C' = pY$$

and by (5) we have

(8) $$E(U) = U(pY)$$

What can we say about the willingness to pay for increases in p when expected utility is as in (4) vs. (8)? If the person with the expectation (4), because insurance is unavailable, can spend dY to obtain dp, he will stop spending when

(9) $$-\frac{dY}{dp} = \frac{U(Y)}{pU'(Y)}$$

I have shown elsewhere (1977) that a person willing to buy insurance has

(10) $$\frac{U(Y)}{U'(Y)} > Y$$

so that

$$(11) \qquad -dY/dp > Y/p$$

In contrast, if insurance is available so that his expectation is (8), he will stop spending when

$$(12) \qquad -dY/dp = Y/p$$

That is, if he is insured, he will be unwilling to pay as high a price to reduce risk as he would be if he is uninsured—the cutoff price given by (12) is below that in (11). (In fact, whereas the uninsured person sacrifices consumption (given good news) to increase safety, the insured person merely pays for safety what he can recover through a reduction in his insurance premium.) The clearly reduced willingness to buy safety contrasts with the uncertain effect found by Ehrlich and Becker; the difference is due to our assumption that safety is paid for only by those who get good news, which differs from their assumption (appropriate for milder casualties) that everyone pays regardless of the news.

Now consider the case where a casualty kills the victim, and where a person facing this risk is willing to buy life insurance if it is available. His utility function is $U_a(\ )$ for good news and $U_d(\ )$ for bad news; we assume that for each $C > C^0$, where C^0 is the starvation level,

$$(13) \qquad U_a(C) > U_d(C) > 0$$

and

$$U_a'(C) > U_d'(C) > 0$$

Where no insurance is available, this person's expected utility is $E(U) = pU_a(Y)$ and so his willingness to buy safety is given in (9), using $U_a(Y)$. One can readily verify that $|dY/dp|$ increases with Y. If fair life insurance is available, the second set of inequalities in (13) implies that the person will insure less than his full lifetime consumption. For small risks, i.e., for $p \rightarrow 1$, we obtain

$$(14) \qquad \lim_{p \rightarrow 1} -\frac{dY}{dp} = C' + \frac{U_a(Y) - U_d(C')}{U_a'(Y)}$$

The reasoning that proved (10) can also be used to show that the expression (14) implies

$$(15) \qquad Y < -\frac{dY}{dp} < \frac{U_a(Y)}{U_a'(Y)}$$

Comparison with (9) shows that in this case also the insured person will spend less for safety than will the uninsured person.

However, the left-hand inequality in (15), together with (9), (11), and (12), show that in every case, whether the person is insured or not, he is willing to pay more than Y as his cutoff price or value of life $-dY/dp$, for p sufficiently close to 1. These cases all involve fair insurance; for these cases, the value of life exceeds the present value of lifetime earnings.

It is also of interest to consider unfair insurance, where

$$(16) \qquad Q > \frac{1-p}{p} C'$$

Let the load ratio L be defined as

$$(17) \qquad L = \frac{pQ}{(1-p)C'} > 1$$

If the person buys life insurance, he chooses C' to obtain $U_d'(C') = LU_a'(Y - Q)$.

The resulting analogue of (15) can be shown to be

$$(18) \qquad Y + (L-1)C' < -\frac{dY}{dp} < \frac{U_a(Y)}{U_a'(Y)}$$

so that although the availability of life insurance lowers the willingness to pay for safety, the load factor implies a higher lower bound for this willingness—lifetime earnings plus the excess load ratio (above one) times the amount insured.

III. Interpretive Comments

Figure 2 shows the value of life implicit in a person's willingness to pay for safety. The ratio $U_a(Y)/U_a'(Y)$ is the quantity $Y + X$, obtained by extending the line tangent to U_a at Y to its intersection with the horizontal axis at $-X$. A person is willing to buy either casualty or life insurance only if $Y > C^*$, where C^* is that level of consumption such that there is a ray from the origin that is tangent to U_a at C^*

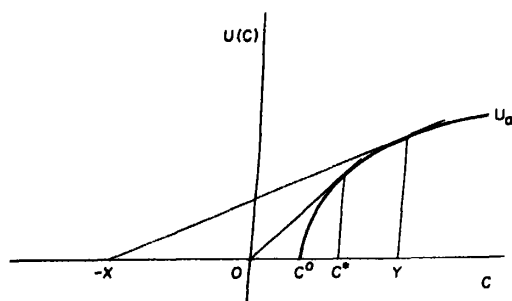



FIGURE 2

(see the author, 1977). Hence, because U_a is concave, $-X$ is to the left of the origin, so that willingness to pay exceeds income as already noted.

A person who buys fair casualty insurance up to full coverage guarantees himself the consumption level $Y - Q$, and so enjoys the benefit of pooling of risks. Hence he is willing to buy reductions of casualty risk only to the point at which the last dollar so spent obtains a dollar reduction in his insurance premium; his "value of preventing total income loss" drops from $Y + X$ to Y/p, or virtually to Y when the risk is small. This "moral hazard" is Pareto optimal, as Ehrlich and Becker pointed out.

However, this Pareto optimality holds only when the insurance premium adjusts exactly to reflect reductions in risk. In contrast, a person fully insured against a casualty will spend *nothing* to reduce casualty risk if the insurance premium remains unchanged, because with full coverage the person maximizes consumption $= Y - Q - S$, where S is his expenditure on risk reduction. With a load factor $L > 1$, and with life insurance, the person is willing to spend something for safety because not all of his risk is insured; but his spending is less than Pareto optimal. This strikes me as showing the practical meaning of moral hazard.

To the extent that insurance premiums fail to adjust for a policyholder's changes in risk, part of these changes are externalities, being shifted to the insurance company's policyholders as a group. Our expressions for the value of safety in the previous section, denoted by $-dY/dp$, would measure the sum of the person's own value plus this externality. (Hence the lower bounds we established there are valid.) To obtain the value of safety accurately, and not simply as a lower bound, one needs to add this externality (plus any others) to the observed willingness to pay.

REFERENCES

Martin J. Bailey, *Measuring the Benefits of Life Saving*, Washington 1978.

———, "Earnings, Life Valuation, and Life Insurance," mimeo. 1977.

G. Blomquist, "Value of Life: Implications of Automobile Seat Belt Use," unpublished doctoral dissertation, Univ. Chicago 1977.

I. Ehrlich and G. S. Becker, "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*, 623–48.

M. Friedman and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *J. Polit. Econ.*, 1948; reprinted in *Readings in Price Theory*, Homewood 1952.

DISCUSSION

ROLAND N. MCKEAN, University of Virginia: Henry Grabowski and John Vernon have presented a useful appraisal of the consequences and prospects of product safety regulations. I agree with their judgments and findings, so my comments will simply qualify, or elaborate on, one of their themes.

Clearly they are correct that agencies do not give much weight to cost-benefit analyses of alternative courses of action, and particularly to costs imposed on consumers. The latter reminds one of the failure of "consumer" groups to get more upset than they do about tariffs, quotas, taxes, price supports, and inflation. This general indifference to impacts on costs suggests that consumer agencies and organizations are soon captured by factions who do not represent the *average* shopper.

And, in a broad sense, it may seem (as the authors indicate) that product-safety agencies follow some sort of "safety imperative"—that our political process treats some degree of safety as a "right" that people should not be allowed to trade for other things, much like the right to vote or to due process. Yet there are many exceptions; government doesn't adopt a safety imperative in any systematic fashion.

For example, as the authors mention, the regulatory agencies could produce *more* safety from their budgets if they tried to be cost effective, that is, to equalize marginal productivity in various life-saving programs. Similarly, agencies could probably allocate effort among enforcement and other activities in such a way as to increase the safety produced with their budgets. (Indeed, efforts that result in unopenable pill containers or safety matchcovers may actually reduce safety.)

More significantly, though, neither the agencies nor Congress will do much about certain extremely hazardous products, such as alcohol or tobacco. Officials are very hesitant to limit the output of nuclear power. With regard to the breeder, they seem willing to accept extra risks of nuclear terrorism. Government could save a huge number of relatively young lives by enforcing lower speed limits on driving. Why does government instead (through the *CPSC*) try to reduce grocery-cart hazards and declare tricycles unsafe at any speed?

To explain these results, I think one must turn to public-choice and bureaucratic-choice models. Along this line, consider the following factors. A significant number of voters, and thence congressmen, and thence administrators, *do* sometimes get concerned with costs. *Ceteris paribus*, the more *direct* the cost of regulations, the *more* likely cost will influence decisions about those regulations. For instance, the voter who wants a drink or a cigarette is acutely aware of inconvenience ahead if those products are threatened. Or, a few years ago, some costs of the automobile interlock system were painfully obvious when people couldn't start their cars unless every package and watermelon was securely belted into its seat. However if grocery carts are made 20 percent more expensive than before, or if toxic substance regulations render many final products more costly, the "tax" on the consumer is indirect and hard to perceive.

Also, if a voter becomes emotionally aroused—enough for him to overcome his inherent free-riderhood—then the cost of a regulation to him personally is *large*. The *bigger* the directly felt cost, the more likely it will be taken into account. A moratorium on nuclear power plants would, as many people see the situation, mean great distress to them personally. They would have to curtail their use of electric hair brushes and other conveniences. Not so with tricycles—many parents could buy their children Hondas instead.

Moreover, voters will overcome their free-rider positions if they become emotionally involved over the *nature* of the issue, that is, if it involves what they see as their "rights." In some instances perhaps they regard the "right to safety" as something that individuals should not be allowed to trade for extra wages or convenience. However when it comes to automobile

driving, human rights to high-speed driving far outweigh such a safety imperative.

Since voters are basically free- or cheap-riders, they will put up with a lot of cost in the aggregate *if* it is highly indirect, *if* it is spread out so as to be small per voter, or *if* the cost isn't an emotionally charged sacrifice. As political and bureaucratic entrepreneurs explore alternatives, we can look forward with confidence to the continued benign neglect of cost-benefit analyses, to the proliferation of often-petty regulations that touch few nerves, and to the disregard of risks where costs of regulation *would* anger numerous voters.

PHILIP J. COOK, Duke University: Martin Bailey's principal result is that "For persons with no assets who buy insurance ... the 'value of life' implicit in safety choices exceeds discounted lifetime earnings." That this result could be demonstrated by a priori reasoning seems surprising. A person who feels that a 50 percent reduction in his income would make his life intolerable (i.e., inferior to death) is not necessarily irrational, yet intuition suggests that such a person is likely to violate Bailey's result. Indeed, a simple algebraic example demonstrates that his result is incorrect.

Using Bailey's notation, we set $U_a(C) = C$ and $U_d(C') = C'^{1/2}$, with budget constraint given by $Y = C + (1-p)\,C'/p$ (assuming fair insurance is available).

These assumptions are compatible with the requirements specified by Bailey in stating his theorem for subsistence income $C^0 > 1$. The specified individual buys life insurance to the point where C' has marginal utility equal to that of lifetime consumption; hence $U_d'(C') = 1/2C'^{-1/2} = U_a'(C) = 1$, so that $C' = .25$. Expected utility is then given by $E(U) = pY + .25\,(1-p)$ (substituting the budget constraint and the solution for C'). It then follows that $-dY/dp = 1/p\ (Y - .25)$, holding $E(U) =$ constant. This of course converges to a number less than Y as $p \rightarrow 1$. This example disproves Bailey's principal result.

As I have shown in a forthcoming paper, under Bailey's assumptions (or slightly more general assumptions) $-p\ dY/dp = Y - C(1 - 1/\alpha) + C'(1 - 1/\beta)$ holding $E(U) =$ constant, where α is the lifetime consumption elasticity of utility, and β is the bequest elasticity of utility, where both elasticities are measured at equilibrium. There are a variety of reasonable circumstances under which this expression is less than Y.

Finally, it should be noted that Bailey measures the value of life in terms of dollars which are paid contingent on the individual's survival. It seems more natural to measure the value of life in terms of "sure" dollars, since investments in safety are costly to society whether or not they are successful in saving the individual's life. This correct expression in terms of sure dollars is equal to the right-hand side of the previous expression, the p is simply eliminated from the expression.

The "Rationality" of Economic Forecasts

By STEPHEN K. MCNEES*

Forecasts can be evaluated relative either to an alternative forecasting procedure (traditionally a naive method) or to a desired statistical property (such as unbiasedness, efficiency, or rationality). This paper employs two statistical tests to examine whether the *ex ante* forecasts of three prominent forecasters were "rational." The first is a test of unbiasedness, a common definition of rationality. Some investigators have taken acceptance of the null hypothesis of unbiasedness as sufficient for inferring rationality, even though unbiasedness need not imply most accurate or best. If rationality is understood in its strong form as incorporating all available information, this test is inadequate. A second test examines whether the forecasts could have been improved on the basis of information available at the time of the forecast. Rather than searching the universe of plausible explanatory information, this investigation is confined to past forecast errors *as they were known to the forecaster at the time the forecast was made*. The final section of the paper discusses the implication of these results as a test of part of the set of propositions known as the rational expectations hypothesis.

I. Were the Forecasts Biased?

Many forecast evaluations have employed Theil's technique of regressing a variable's "realization" (i.e., its actual value A) on its forecasted (or predicted) value P.

$$A = \alpha + \beta P + u \qquad (1)$$

An F test can be employed to examine the joint null hypothesis that $\alpha = 0$ and $\beta = 1$, the values consistent with unbiased forecasts.

Estimation of (1) requires some assumptions about the probability structure of the disturbances (u). Typically, it is assumed that the disturbances are serially uncorrelated and the ordinary least squares (*OLS*) technique is applied. This assumption is inappropriate for multiperiod forecasts issued at quarterly intervals where serial dependence is a by-product of defining multiperiod realizations as cumulative changes during overlapping time intervals. A random shock (or a measurement error) occurring in one period which is not subsequently offset will appear in n consecutive A_{t+n}. If it is unanticipated (not reflected in the P's) and is not reversed by an offsetting shock, it will also appear in n consecutive u's. Under these assumptions, the variance-covariance matrix can be specified precisely and the coefficients of equation (1) estimated more efficiently by generalized least squares (*GLS*). Because these assumptions may not always hold precisely, both the *GLS* and *OLS* results are presented below.

Three forecasters' one-, two-, three-, and four-quarter ahead forecasts of the implicit *GNP* price deflator (*IPD*), real *GNP*, and the unemployment rate (*UR*) were tested.[1] Using the *OLS* results in Table 1, the null hypothesis is accepted for fourteen of the

*Assistant vice president and economist, Federal Reserve Bank of Boston. I am indebted to colleagues at the Boston Fed, especially Richard Kopcke and Donald Rindler, for invaluable assistance. Otto Eckstein, Michael K. Evans, and Lawrence R. Klein generously provided both forecast data and encouragement. Ray Fair, Benjamin Friedman, and William Poole provided helpful comments. I bear sole responsibility for errors of fact or interpretation.

[1]The three forecasters selected were those most meticulous and consistent in terms of the timing of their release dates. They are also among the most frequently cited and widely used.

TABLE 1—TEST RESULTS: *OLS* ESTIMATES

Forecast Horizon in Quarters	Chase				*DRI*				Wharton			
	α	β	F	γ	α	β	F	γ	α	β	F	γ
					Implicit *GNP* Price Deflator (*IPD*) (Percent change, annual rate)							
1	.33	1.04	1.00	0.6[b]	.51	1.10	2.24	0.5[b]	−.20	1.12	2.36	0.1
	(.99)	(.16)		(0.2)	(1.38)	(.24)		(0.2)	(.68)	(.11)		(0.2)
2	1.32	.93	1.84	0.4	1.37	.99	3.48	0.4	.62	1.08	4.16[b]	0.2
	(1.27)	(.21)		(0.2)	(1.38)	(.25)		(0.2)	(.89)	(.15)		(0.2)
3	1.63	.92	2.69	0.3	1.65	.98	4.39[b]	0.4	.99	1.06	4.91[b]	0.4
	(1.45)	(.26)		(0.3)	(1.43)	(.27)		(0.3)	(1.03)	(.18)		(0.2)
4	2.32	.83	3.18	0.2	2.25	.91	5.06[b]	0.3	1.61	.98	4.49[b]	0.2
	(1.70)	(.31)		(0.3)	(1.54)	(.30)		(0.3)	(1.27)	(.23)		(0.3)
					Real Gross National Product (real *GNP*) (Percent change, annual rate)							
1	−1.61	1.35	4.94[b]	0.5[b]	−2.30	1.41	4.83[b]	0.6[b]	−1.84	1.34	3.96[b]	0.4
	(.62)	(.12)		(0.2)	(.79)	(.16)		(0.2)	(.72)	(.14)		(0.3)
2	−2.36	1.42	5.11[b]	0.1	−4.01	1.66	7.79[a]	0.4	−2.73	1.42	5.47[b]	0.2
	(.75)	(.16)		(0.2)	(1.01)	(.21)		(0.2)	(.76)	(.16)		(0.2)
3	−3.08	1.56	4.64[b]	0.3	−4.31	1.73	4.71[b]	0.3	−3.31	1.61	6.50[a]	0.3
	(1.01)	(.24)		(0.3)	(1.46)	(.34)		(0.3)	(.84)	(.19)		(0.3)
4	−4.31	1.80	6.54[a]	0.6	−5.30	1.88	5.54[b]	0.6	−3.79	1.67	7.56[a]	0.5
	(1.22)	(.30)		(0.4)	(1.77)	(.41)		(0.4)	(.90)	(.21)		(0.4)
					Unemployment Rate (*UR*) (Percentage points, cumulative change)							
1	.08	1.13	1.09	0.2	.01	1.06	.29	0.0	−.03	1.16	.39	0.2
	(.06)	(.19)		(0.2)	(.06)	(.15)		(0.2)	(.07)	(.18)		(0.2)
2	.16	1.33	2.08	−0.1	.07	1.23	1.33	−0.2	.10	1.32	1.37	−0.1
	(.12)	(.23)		(0.2)	(.11)	(.17)		(0.2)	(.18)	(.25)		(0.2)
3	.28	1.56	3.97[b]	−0.1	.13	1.44	3.51[b]	−0.1	.28	1.36	2.37	−0.1
	(.16)	(.27)		(0.3)	(.14)	(.20)		(0.3)	(.17)	(.27)		(0.3)
4	.42	1.86	6.75[a]	0.8	.21	1.56	4.16[b]	0.4	.50	1.45	4.71[b]	0.4
	(.19)	(.31)		(0.8)	(.19)	(.24)		(0.8)	(.20)	(.26)		(0.4)

Notes: Sample period: 1970:II to 1975:IV for *IPD* and real *GNP*; 1970:II to 1976:II for *UR*. The standard errors are shown in parentheses. The *F*-statistic refers to a test of the joint hypotheses, $\alpha = 0, \beta = 1$.
[a]Significant at the 1 percent level.
[b]Significant at the 5 percent level.

thirty-six regressions (3 forecasters × 3 variables × 4 horizons). Using *GLS* for the multiperiod forecasts in Table 2, the null hypothesis is accepted in twenty of the thirty-six cases.

The null hypothesis cannot be accepted for the regressions taken as a whole, treating each as if it were independent.[2] Under this independence assumption, the results for each regression can be considered as elementary events constituting a set of random experiments with two mutually exclusive and exhaustive outcomes—acceptance or rejection of the null hypothesis. In this manner, it is possible to test the general hypothesis that forecasts *as a whole* are unbiased. Application of a binomial test, based on the 5 percent level of significance for each regression, demonstrates that the number of rejections is well in excess of

[2]The results are clearly not independent. However, it is impossible to determine whether this raises or lowers the critical boundary without knowing more about the structure of the dependencies.

TABLE 2—TEST RESULTS: *GLS* ESTIMATES

Forecast Horizon in Quarters	Chase				*DRI*				Wharton			
	α	β	F	γ	α	β	F	γ	α	β	F	γ
					Implicit *GNP* Price Deflator (*IPD*) (Percent change, annual rate)							
2	5.20	.20	17.38[a]	0.7[a]	5.19	.21	12.50[a]	0.6[a]	4.70	.29	12.08[a]	0.6[a]
	(1.05)	(.14)		(0.2)	(1.12)	(.16)		(0.2)	(1.05)	(.15)		(0.2)
3	4.05	.44	5.67[b]	0.2	4.88	.32	7.54[a]	0.4	4.39	.40	9.24[a]	0.6[a]
	(1.20)	(.19)		(0.2)	(1.26)	(.22)		(0.4)	(1.04)	(.16)		(0.2)
4	5.76	.13	14.70[a]	0.4	4.87	.30	5.60[b]	0.1	4.21	.44	5.65[b]	0.1
	(1.18)	(.17)		(0.3)	(1.46)	(.24)		(0.3)	(1.25)	(.20)		(0.3)
					Real Gross National Product (real *GNP*) (Percent change, annual rate)							
2	−1.50	1.13	.93	0.1	−.36	.73	1.37	0.5[b]	−.69	.85	.97	0.3
	(1.13)	(.21)		(0.2)	(1.45)	(.24)		(0.2)	(1.27)	(.21)		(0.2)
3	.07	.65	1.43	0.3	.27	.53	2.15	0.3	.98	.41	3.46	0.6[a]
	(1.53)	(.25)		(0.3)	(1.76)	(.28)		(0.3)	(1.58)	(.24)		(0.2)
4	.76	.51	1.03	0.2	1.95	.12	2.36	0.2	1.10	.39	1.60	0.6
	(1.81)	(.37)		(0.5)	(2.07)	(.43)		(0.5)	(1.84)	(.36)		(0.4)
					Unemployment Rate (*UR*) (Percentage points, cumulative change)							
2	.21	.67	1.66	0.1	.13	.77	.88	0.2	.18	.36	6.34[a]	0.5[a]
	(.18)	(.23)		(0.2)	(.16)	(.20)		(0.2)	(.11)	(.18)		(0.2)
3	.30	.64	1.38	0.1	.26	.51	2.41	0.3	.31	.67	1.11	−0.0
	(.28)	(.27)		(0.2)	(.29)	(.23)		(0.2)	(.29)	(.31)		(0.2)
4	.48	.40	4.23[b]	0.1	.46	.31	7.75[a]	0.4	.47	−.12	14.89[a]	0.5[b]
	(.40)	(.23)		(0.3)	(.40)	(.18)		(0.3)	(.43)	(.17)		(0.2)

Notes. See Table 1.
[a]Significant at the 1 percent level.
[b]Significant at the 5 percent level.

what would be expected to occur by chance alone if the null hypothesis held for each individual regression. This result holds for both the *GLS* and the *OLS* estimates, for all regressions, and for each variable taken separately. The findings are very similar among forecasters. It would be wise, therefore, to resist the temptation to single out any specific forecaster's predictions as biased or unbiased, rational or irrational, as compared to the others.

Except for the two preceding conclusions, the results are quite sensitive to the estimation technique employed. Using *OLS*, the null hypothesis is rejected for all real *GNP* forecasts and most multiperiod *IPD* forecasts. Using *GLS*, the null hypothesis is accepted for all multiperiod real *GNP* forecasts and is rejected for all multiperiod *IPD* forecasts. The overall results for *UR* are not very sensitive to the choice of estimation technique.

The results are highly dependent on the horizon of the forecast. For example, although the null hypothesis is accepted for one-quarter ahead *IPD* forecasts, it is either consistently (with *GLS*) or typically (using *OLS*) rejected for multiperiod forecasts of *IPD*. Although the null hypothesis is not accepted for one-quarter ahead real *GNP* forecasts, it is consistently accepted for multiperiod forecasts when *GLS* is employed. The null hypothesis is accepted for one-quarter ahead *UR* forecasts and re-

jected for the four-quarter ahead horizon.

Finally, unbiasedness (i.e., acceptance of the null hypothesis) need not imply best (in the sense of minimum mean absolute error or root-mean squared error). This calls into question the power of this F test to assess the rationality of forecasts. For example, consider a forecaster whose predictions always missed the mark by $\pm x$ but whose underestimates and overestimates were equally likely. If x were large, these forecasts would be highly inaccurate, or inefficient. There would be good reason to question whether these forecasts had incorporated all information available at the time the forecast was made, a stronger form of rationality.

II. Did the Forecasts Incorporate all Information?

Incorporation of all available information implies that forecast errors should not be systematically related to any information known at the time the forecast was made. It is beyond the scope of this paper to construct a model that attempts to explore alternative structural explanations of the relevant information set. Instead, the investigation is confined to forecast errors. These errors, ϵ^*, are defined as the forecast errors as known by the forecaster at the time of the forecast. They are based on the preliminary data available to the forecaster and are not therefore simply lagged forecast errors which are based on the revised data. As a second test, therefore, equation (2) is estimated,

$$\epsilon_{t+n} = \gamma \epsilon_{t-n} + \nu_{t+n} \quad (2)$$

and the hypothesis that $\gamma = 0$ is tested. Of course, limiting the investigation to the subset of information contained in the forecaster's previous errors is a weak test but failure to pass such a test indicates that readily available information was ignored and must surely call into doubt the rationality of the forecasts.

As shown in Table 1, when (2) is estimated by OLS, only three of the thirty-six γ's are significantly different from zero at the 5 percent level. Application of the binomial test shows that this number of rejections of the hypothesis $\gamma = 0$, could have occurred by chance alone. However, all three of the significant coefficients appear in the nine one-quarter ahead regressions and this is well above the number that could be expected to occur by chance alone. As noted above, multiperiod forecasts covering overlapping time intervals exhibit serial dependence which, in general, renders OLS inappropriate.

Using the GLS results for multiperiod forecasts shown in Table 2, the past forecast error was not significantly related to the *ex ante* forecast error in twenty-five of the thirty-six cases. In the remaining eleven cases, significant coefficients were found, well above the number that could be expected to occur by chance alone if the null hypothesis held for each regression. This test, then, indicates rejection of the general hypothesis that forecasts *as a whole* cannot benefit from the information contained in past errors. The same conclusion holds for both the one-quarter ahead forecasts (estimated by OLS) and the multiperiod forecasts (estimated by GLS) taken separately. It also holds for IPD and for real GNP taken separately, but not for UR alone.

Again the results among different forecasters show a general similarity, although it is less striking than for the first test. In half of the cases, all forecasters were alike in either benefiting (two-quarter ahead IPD only) or not benefiting from past errors. Two forecasters could benefit in three cases and one benefits in five cases.

III. Summary and an Alternative Interpretation

Using the GLS estimates, the two null hypotheses of rationality—unbiasedness and minimum squared error—are accepted by both tests in seventeen of the thirty-six cases and rejected by both in eight. In the remaining eleven cases, the results of the two tests differ: the forecasts were biased but could not have been improved with the information in past forecast errors eight

times, and unbiased but significantly related to past errors three times. These results clearly indicate that these forecasts were not rational over this forecast period.

These results can also be interpreted as evidence bearing on one version of part of the set of propositions known as the rational expectations hypothesis (*REH*). Previous empirical investigations of the *REH* have been criticized for two reasons: (a) several *jointly* test rationality along with some other hypothesis so the results can be taken as tests of rationality only if the jointly tested model were assumed true; and (b) all employ various proxies as measures of expectations with little assurance that actual decisions were based on this information. In contrast, this study has measured expectations directly as the *ex ante* forecasts by the most prominent forecasters. These forecasts have met the market test—they were purchased by the policymaking agencies of government and business. Although it could be argued that behavior is based on rationally adjusted versions of these forecasts, to do so seems unwarranted without further documentation. Commercial forecasters are acknowledgedly skillful processors of economic information with a close link between their rationality and payoff.

Whatever interpretation is given to these results, their major limitation is the short, possibly unrepresentative, sample period on which they are based. It has been argued elsewhere that the forecast period is the major source of variation in forecast accuracy (see the author). It is generally agreed that the early 1970's were a particularly tumultuous period when economic events were strongly influenced by "noneconomic" factors. It is ironic that confidence in forecasting seemed higher in the 1960's even though forecasting *techniques* may have improved since then and clearly have not deteriorated.

REFERENCE

S. K. McNees, "The Forecasting Performance in the 1970s," Federal Reserve Bank of Boston, (an updated and corrected version of "The Forecasting Performance in the Early 1970s," *New England Econ. Rev.*, July/Aug. 1976, 1-13).

An Error Analysis of Econometric and Noneconometric Forecasts

By VINCENT SU*

*In this study, the historical records of the Wharton quarterly forecast and the ASA-NBER survey forecast are studied. The complete record of the Wharton forecast in-*cludes at least two sets of forecasts for each quarter. Each forecast projects eight-ten quarters ahead. The premeeting forecast is usually made at the end of the first month of each quarter, two weeks before the Wharton quarterly meeting of its users. This forecast is made based on the preliminary data released fifteen days after the end of the quarter. The postmeeting forecast is usually made at the end of the second month of each quarter and is based on the forty-five day release data. In addition to the difference in the data, the postmeeting forecast contains more information that has become available in the second month and also includes the feedback from the Wharton users in industry and government.

The *ASA-NBER* survey forecast is a quarterly consensus forecast which was first released in December 1968. In each quarter, the questionnaires are sent to a regular panel of roughly 160 economists. Each is asked to predict ten major economic variables for four or five quarters ahead. Usually, about 40 to 80 of the panel members return the questionnaires. On average, the majority (57 percent) of the panel used an informal *GNP* model approach as the most important technique for making their forecast; 25 percent of them used econometric models, 9 percent used leading indicators, and 8 percent used other unspecified methods. The users of econometric models increased significantly in 1972 and 1973; however, there was a slight decrease in the 1974–75 period. The users of leading indicators declined from 15–20 percent in 1969 to 2–3 percent in 1976 (see the author and Josephine Su for a detailed description).

The 1976 national income accounts data revision has been a difficult problem to handle in error assessment. In addition to the change of base year for constant dollar accounts, this revision also included some definitional modification in *GNP* and its components. This modification causes an inconsistency in forecast series. The best way to avoid this difficulty is to convert both forecast series and actual data into percentage changes or cumulative changes. The forecasting error is then measured as the difference between the forecast percentage change and the actual percentage change.

I

Tables 1 and 2 contain calculations of the root-mean square errors (*RMSE*) of the Wharton premeeting and postmeeting forecasts and the *ASA-NBER* forecasts for ten *ASA-NBER* variables plus real *GNP*. Since the Wharton model did not produce forecasts of plant and equipment expenditures calculated from the Bureau of Economic Analysis (*BEA*) survey until very recently, the fixed nonresidential investment in national income accounts is used as a substitute. The forecast of the industrial production index was appended to the Wharton model in 1977, therefore, its *RMSE* are not included.

The *RMSE* are for the percentage changes and cumulative percentage changes of the three sets of forecasts, except for *UR* and *II$* which were calculated from the quarter-to-quarter changes and the cumulative changes. The actual series used in calculating *RMSE* are the latest available data. The sample period is from 1968IV to

*Baruch College, City University of New York. I am indebted to Lawrence Klein and Josephine Su for their helpful suggestions and comments.

TABLE 1—THE COMPARISON OF *RMSE* OF *ASA/NBER* AND WHARTON FORECASTS, 1968IV–1977II

	RMSE of Predicted Percentage Changes							
	1Q	2Q	3Q	4Q	5Q	6Q	7Q	8Q
Variable *GNP$* (Gross national product, current billion $)								
ASA/NBER	0.65	0.83	0.92	0.93	—	—	—	—
Wharton Premeeting	0.74	0.93	0.89	0.93	1.01	1.07	1.07	1.11
Wharton Postmeeting	0.71	1.07	0.94	0.92	1.02	1.01	1.10	1.10
Variable *GNP* (Gross national product, constant billion $)								
ASA/NBER	0.66	0.88	1.10	1.19	—	—	—	—
Wharton Premeeting	0.67	0.94	0.96	1.10	1.21	1.17	1.18	1.26
Wharton Postmeeting	0.69	1.04	0.95	1.07	1.14	1.16	1.20	1.23
Variable *IPD* (*GNP* implicit price deflator, 1972 = 100)								
ASA/NBER	0.40	0.56	0.71	0.76	—	—	—	—
Wharton Premeeting	0.33	0.52	0.67	0.71	0.81	0.87	0.89	0.90
Wharton Postmeeting	0.34	0.44	0.64	0.69	0.73	0.85	0.88	0.93
Variable *CD$* (Consumer expenditures for durables, current billion $)								
ASA/NBER	2.38	3.44	3.54	3.51	—	—	—	—
Wharton Premeeting	2.51	3.18	3.72	3.58	3.73	3.75	3.64	3.99
Wharton Postmeeting	2.46	3.29	3.78	3.74	3.91	3.72	3.81	3.95
Variable *PE* (Plant and equipment expenditures, current billion $)								
ASA/NBER	1.86	2.00	2.03	2.11	—	—	—	—
Wharton Premeeting	1.68	2.32	2.96	2.34	3.31	2.88	2.75	2.64
Wharton Postmeeting	1.88	2.31	2.32	2.36	2.34	2.73	2.86	2.78
Variable *II$* (Change in business inventories, current billion $)								
ASA/NBER	7.94	9.58	9.94	10.12	—	—	—	—
Wharton Premeeting	7.69	10.10	10.35	10.16	10.92	10.70	10.90	10.95
Wharton Postmeeting	7.69	10.86	10.16	10.44	10.41	10.85	10.95	10.99
Variable *NDP$* (National defense purchases, current billion $)								
ASA/NBER	1.86	1.88	1.94	2.01	—	—	—	—
Wharton Premeeting	1.84	1.77	2.01	2.05	1.87	1.92	1.62	1.70
Wharton Postmeeting	1.74	1.63	2.07	1.91	1.97	1.82	1.68	1.58
Variable *CPAT$* (Corporate profits after taxes, current billion $)								
ASA/NBER	5.99	6.31	7.55	7.62	—	—	—	—
Wharton Premeeting	7.21	7.33	7.56	8.04	8.53	8.21	7.88	8.55
Wharton Postmeeting	6.45	7.88	7.58	7.76	8.32	8.88	8.33	7.78
Variable *UR* (Unemployment rate, percent)								
ASA/NBER	0.17	0.33	0.39	0.44	—	—	—	—
Wharton Premeeting	0.28	0.39	0.39	0.42	0.45	0.48	0.46	0.49
Wharton Postmeeting	0.22	0.35	0.44	0.42	0.42	0.45	0.45	0.49
Variable *HS* (New private housing units started, annual rate million)								
ASA/NBER	6.06	9.71	8.45	5.99	—	—	—	—
Wharton Premeeting	10.68	10.65	9.35	9.80	8.96	7.22	7.48	6.59
Wharton Postmeeting	7.01	9.36	7.94	7.87	9.61	6.81	7.03	8.62
Variable *IP* (Industrial production index, 1967 = 100)								
ASA/NBER	1.51	2.03	2.41	2.49	—	—	—	—

1977II, except for *NDP$*, *CPAT$*, and *HS* which begin at 1971II, 1969III and 1974II, respectively. Early versions of the Wharton model did not forecast these variables.

It is obvious from Tables 1 and 2 that the accuracy of all forecasts deteriorates as the forecasting horizon expands. The accuracy deteriorates more rapidly in the first two quarters. In fact, the *RMSE* of percentage changes are rather stable after the second quarter for most variables. The *RMSE* of aggregate variables such as *GNP$*, *GNP*, *IPD*, and *UR* are relatively small. The major *GNP* components, i.e., *CD$*, *PE$*, and *NDP$*, have larger *RMSE*, and the largest *RMSE* are found with *CPAT$* and *HS*.

The *ASA-NBER* survey outperforms the Wharton model in predicting six variables in the first quarter. The superiority of the

TABLE 2—THE COMPARISON OF *RMSE* OF *ASA/NBER* AND WHARTON FORECASTS, 1968IV–1977II

	RMSE of Predicted Cumulative Percentage Changes							
	1Q	2Q	3Q	4Q	5Q	6Q	7Q	8Q
Variable *GNP$* (Gross national product, current billion $)								
ASA/NBER	0.65	1.10	1.41	1.65	—	—	—	—
Wharton Premeeting	0.74	1.24	1.57	1.77	2.16	2.42	2.74	3.26
Wharton Postmeeting	0.71	1.34	1.75	1.91	2.17	2.37	2.66	3.20
Variable *GNP* (Gross national product, constant billion $)								
ASA/NBER	0.66	1.11	1.66	2.33	—	—	—	—
Wharton Premeeting	0.67	1.13	1.48	1.97	2.59	3.14	3.69	4.29
Wharton Postmeeting	0.69	1.28	1.62	2.02	2.46	2.93	3.52	4.17
Variable *IPD* (*GNP* implicit price deflator, 1972 = 100)								
ASA/NBER	0.40	0.89	1.55	2.31	—	—	—	—
Wharton Premeeting	0.33	0.78	1.36	2.04	2.85	3.75	4.67	5.65
Wharton Postmeeting	0.34	0.69	1.27	1.91	2.62	3.50	4.42	5.41
Variable *CD$* (Consumer expenditures for durables, current billion $)								
ASA/NBER	2.38	4.12	5.23	6.09	—	—	—	—
Wharton Premeeting	2.51	4.41	5.76	6.78	7.73	8.75	10.03	11.33
Wharton Postmeeting	2.46	4.95	6.53	7.46	8.16	9.24	10.53	11.98
Variable *PE$* (Plant and equipment expenditures, current billion $)								
ASA/NBER	1.86	3.08	3.91	4.98	—	—	—	—
Wharton Premeeting	1.68	3.09	4.98	6.30	7.29	9.17	11.13	13.12
Wharton Postmeeting	1.88	3.42	4.96	6.42	7.96	9.60	11.75	13.85
Variable *II$* (Change in business inventories, current billion $)								
ASA/NBER	7.94	8.77	9.58	10.06	—	—	—	—
Wharton Premeeting	7.69	10.05	11.06	11.32	13.37	14.62	13.76	13.68
Wharton Postmeeting	7.69	9.20	11.58	11.00	12.84	13.77	13.20	12.71
Variable *NDP$* (National defense purchases, current billion $)								
ASA/NBER	1.86	2.73	2.91	3.07	—	—	—	—
Wharton Premeeting	1.84	2.59	2.83	3.52	4.36	5.05	6.18	5.79
Wharton Postmeeting	1.74	2.42	2.46	2.65	2.74	3.35	3.64	3.97
Variable *CPAT$* (Corporate profits after taxes, current billion $)								
ASA/NBER	5.99	9.42	11.81	14.02	—	—	—	—
Wharton Premeeting	7.21	9.59	11.29	12.68	15.32	17.72	20.23	24.50
Wharton Postmeeting	6.45	10.24	11.84	13.28	13.90	13.33	15.34	19.42
Variable *UR* (Unemployment rate, percent)								
ASA/NBER	0.17	0.41	0.66	0.89	—	—	—	—
Wharton Premeeting	0.28	0.55	0.78	0.97	1.08	1.18	1.32	1.56
Wharton Postmeeting	0.22	0.46	0.77	0.94	1.03	1.12	1.24	1.43
Variable *HS* (New Private housing started, annual rate million)								
ASA/NBER	6.06	13.27	19.67	23.70	—	—	—	—
Wharton Premeeting	10.68	16.83	20.40	23.15	26.45	28.12	32.07	37.48
Wharton Postmeeting	7.01	13.62	18.57	22.18	27.41	30.08	32.77	36.90
Variable *IP* (Industrial production index, 1967 = 100)								
ASA/NBER	1.51	2.92	4.00	4.83	—	—	—	—

ASA-NBER survey, however, declines very rapidly as the forecast span extends. By the fourth quarter, the *ASA-NBER* survey still forecast better in four variables, but not significantly better, with the exception of *HS*. The conclusion is that the Wharton forecast becomes relatively better as the forecast period expands. The relative inferiority of the Wharton model in the early quarters could be a consequence of overuse of fine tuning procedures, such as adjusting constant terms. Model builders customarily use these procedures to fine tune their near-term forecasts according to extraneous information about the future. The excessive fine tuning of some variables may distort the prediction of some other variables.

From the variable by variable comparison, it is obvious that the *ASA-NBER* survey is consistently better than the Wharton model in predicting *CPAT$* and

UR. The Wharton model is superior to the *ASA-NBER* survey in forecasting *IPD* and *NDP$*. It is interesting to note in this regard that *NDP$* is an exogenous variable in the Wharton model, its forecast is made from budgetary information and judgment. For other variables, it is difficult to draw a firm conclusion.

As shown in Table 2, the *RMSE* of cumulative percentage changes increase for all forecasts over the eight-quarter period, but the increment decreases after the second quarter. The *RMSE* of four-quarter cumulative percentage changes indicates that the *ASA-NBER* forecast has the smaller error in five variables: *GNP$*, *CD$*, *PE$*, *II$*, and *UR*. The Wharton model is superior in predicting the cumulative changes in *GNP*, *IPD*, *NDP$*, *CPAT$*, and *HS*. The *RMSE* of cumulative percentage change of all eight quarters indicate that the postmeeting Wharton forecast is better than the premeeting forecast in all variables except *CD$* and *PE$*.

Generally speaking, since the middle of 1973 the *U.S.* economy has been in the midst of a series of setbacks caused by some new exogenous factors. Since forecasting with an econometric model is in essence an extrapolation beyond the sample period over which the model was constructed, the model should have difficulty forecasting in this period even with the use of fine tuning procedures. On the other hand, it is relatively easy for the *ASA-NBER* members to adapt their prediction to drastic exogenous shocks since they do not necessarily conform to a structural system. It is interesting to compare the predictive performance of these two forecasting systems over the different stages of the business cycles.

The sample period used in this study, 1968IV–1977II, includes two complete business cycles. The first cycle is a moderate one with a peak at 1969IV and trough at 1970IV. The second cycle is a vigorous one, with a peak at 1973IV and trough at 1975II. The sample period is therefore broken into three subsamples as follows: 1968IV–1973II; 1973III–1975II; and 1975III–1977II.

The results show that the *RMSE* of most variables in the second period are substantially worse than in the first period. This finding prevails in both *ASA-NBER* and Wharton forecasts. Large *RMSE* are found in *CD$*, *II$*, *CPAT$*, *HS*, and *IP* in the second subsample. This suggests that these variables are difficult to predict when the economy is in a severe recession. Surprisingly, the *RMSE* of the near-term Wharton forecast of *IPD* are very close in the first and second periods even though the period 1973III–75II was one of very high inflation. It is also interesting to find that the only variable which was forecast slightly better in the severe recession by both systems is *NDP$*, a lagging indicator by nature. As the war in Vietnam wound down, it is possible that defense expenditures become more predictable.

When the economy swings upward, the *RMSE* of most variables are improved. Only the near-term predictions of *II$* and *UR* are worse than when the economy was in a recession. This is probably because high unemployment is not usually expected in a recovery period. High inflation persisted in this period, but the predictions of *IPD* were substantially improved in all forecasts.

In general, the results for the three periods are mixed. Over the period of the moderate business cycle, the Wharton forecasts of *GNP*, *IPD*, *CD$*, and *PE$* were better, while the *ASA-NBER* forecasts of *CPAT$* and *UR* were consistently better. Thus, the Wharton model may be said to retain a slight edge over the *ASA-NBER* forecasts. The Wharton forecasts were worse than the consensus forecasts in the severe recession period. Only *IPD* is forecast consistently better by Wharton in this period. In the period of a slow recovery, the comparison is, however, inconclusive.

II

This section analyzes forecast inaccuracies caused by the data revisions. A forecast made in the first month of the t-th quarter, F_t^1, is based on the preliminary

data set of the t−1 quarter released on the fifteenth day of the month; these data are called the fifteenth-day release (P_{t-1}^{15}). This data set will be revised after a month. The revised data are called the forty-five-day release (P_{t-1}^{45}). Most forecasters also modify their forecast to take into account the revisions. The modified forecast F_t^2 is, of course, based on the revised data, P_{t-1}^{45}. In the third month of the t-th quarter, the data set will be again revised, this revision is called the seventy-five-day release (P_{t-1}^{75}). After the seventy-five-day release, usually, there are two or more annual revisions before the data are finalized. The final data are assumed to be the actual (A_{t-1}) (see Allan Young for a detailed description).

As time advances into the t+1 quarter, the data of the t-th quarter become available. The data of the t-th quarter are released and revised in the same procedure as described above. The fifteen-day release (P_t^{15}) is the first available estimate of A_t. Had the forecasters possessed as much information as the source agency in the t+1 quarter, they would probably have made their forecast equal to P_t^{15}. Therefore, the difference between the forecast and preliminary data is a part of the error made by the forecasters. When P_t^{45} becomes available in the second month of the t+1 quarter, it is regarded as a better estimate of A_t. The difference between P_t^{15} and P_t^{45} is the part of the revision error solely resulting from the inaccuracy of P_t^{15}. The difference between P_t^{45} and A_t is also a component of revision error due to the remaining inaccuracies associated with P_t^{45}.

We have only decomposed the mean square errors (*MSE*) of the first quarter forecast, because the first-quarter forecast is not affected by the cumulative error that arises through the use of lagged variables. Both forecasts and actuals are measured in percentage changes. The percentage change of preliminary data P_t^{15} and P_t^{45} are calculated as $[(P_t^{15}/P_{t-1}^{75}) - 1]$ and $[(P_t^{45}/P_{t-1}^{75}) - 1]$, since the seventy-five-day release of the t−1 quarter is used as the jump off for calculating P_t^{15} and P_t^{45}. The sample period used is 1968IV to 1974IV, because the latest available data after 1974IV are not yet final. Only eight variables are studied: *II$* is not included because it is originally measured in changes; *HS* and *IP* are dropped because of lack of appearance in former Wharton forecasts.

The formula used to decompose the *MSE* of the *ASA-NBER* survey and the Wharton premeeting forecast is as follows:

$$\sum (F_t^1 - A_t)^2/N = \sum (F_t^1 - P_t^{15})^2/N + \sum (P_t^{15} - A_t)^2/N + 2\sum (F_t^1 - P_t^{15})(P_t^{15} - A_t)/N$$

The formula used to decompose the *MSE* of the Wharton postmeeting forecast is as follows:

$$\sum (F_t^2 - A_t)^2/N = \sum (F_t^2 - P_t^{45})^2/N + \sum (P_t^{45} - A_t)^2/N + 2\sum (F_t^2 - P_t^{45})(P_t^{45} - A_t)/N$$

The results of decomposition of *MSE* into forecast errors and revision errors are reported in Table 3. The first column of each forecast is the *MSE* with respect to the finalized data. The second columns of the *ASA-NBER* and Wharton premeeting forecasts are the *MSE* with respect to the fifteen-day release, the second column of the Wharton postmeeting forecast is the *MSE* with respect to the forty-five-day release data. The revision error (*RE*) of the *ASA-NBER* and Wharton premeeting forecasts is the *MSE* measured as the difference between the fifteen-day release and the actual, while the *RE* of the Wharton postmeeting forecast is the *MSE* measured as the difference between the forty-five-day release and the actual. The last column of each forecast is, of course, the covariance between the forecast error (*FE*) and *RE*.

In Table 2, total errors are always less than the sum of *FE* and *RE*, hence all covariances are negative. In other words, the forecast and the revision generally have a tendency to move in opposite directions. A comparison of the first and second columns of each forecast indicates that the

TABLE 3—THE DECOMPOSITION OF *MSE* OF *ASA-NBER* AND WHARTON FORECASTS 1968IV–1974IV

	ASA-NBER				Wharton Premeeting				Wharton Postmeeting			
	Total	*FE*	*RE*	Cov	Total	*FE*	*RE*	Cov	Total	*FE*	*RE*	Cov
GNP$	.2693	.1971	.1941	−.1192	.3723	.2265	.1914	−.0156	.3275	.1986	.1624	−.0317
GNP	.2667	.2992	.1914	−.2239	.3106	.3153	.1914	−.1961	.2729	.1873	.1914	−.1058
IPD	.2095	.3000	.0748	−.1653	.1183	.1647	.0748	−.1212	.1302	.1680	.0770	−.1148
CD$	4.93	9.14	11.96	−16.17	4.24	8.25	11.96	−15.97	3.66	7.58	11.35	−15.27
PE$	4.21	5.02	2.48	−3.29	3.36	4.59	2.48	−3.70	4.12	5.50	2.13	−3.52
NDP$	3.99	4.61	4.62	−5.24	4.14	6.00	4.62	−6.48	3.77	7.72	7.08	−11.02
CPAT$	25.44	34.08	29.92	−38.57	58.46	63.71	29.92	−35.17	41.83	47.43	25.30	−30.89
UR	.0196	.0144	.0103	−.0051	.0529	.0529	.0103	−.0103	.0350	.0256	.0103	−.0009

*FE*s with respect to the preliminary data are greater than the *FE* with respect to the final data, except *GNP$*, *UR*, and the real *GNP* in Wharton postmeeting forecast. This means that the forecasters are predicting the final values better than the preliminary data, although the forecasts are made on preliminary data. On comparing the predictive power of the forecasts with that of preliminary data, it is found that the forecast of *CD$* and *NDP$* made by any one of the three forecasts are closer to the final data than are the corresponding preliminary data.

Generally speaking, and as might be expected, the forty-five-day release is a better estimate than the fifteen-day release, being closer to the final data. Among the variables studied, the differences between these two sets of preliminary data are not very large. The relatively large differences are found in *GNP$*, *PE$*, and *CPAT$*. The only variable which has a larger error in the forty-five-day release than in the fifteen-day release is *NDP$*. The reason for this is unclear.

III

The ability to identify turning points in business cycles before they occur is an important criterion for evaluating a forecasting system. In general, the reference cycles can be well represented by the cyclical movements in the real *GNP* series. In the sample period used in this study, there are six turning points in the real *GNP* series. These are 1969IV; 1970II; 1970IV; 1971I; 1974I; 1975II.

In general, the *ASA-NBER* survey performs poorly in giving early warning of turning points. It was able to predict only the 1974I peak one-quarter ahead, the change of direction of *GNP* growth was signaled in the forecast made in 1973IV. The Wharton model picked up the 1974I turn in advance, and the Wharton premeeting forecast was able to predict the 1969IV peak one-quarter ahead. It also predicted the upturn in 1975II three quarters before it happened. Nevertheless, subsequent forecasts missed this upturn by plus or minus one quarter. The Wharton postmeeting forecast is most capable of capturing turning points. It gave one-quarter ahead warning of the three important turning points in 1969IV, 1971I, and 1974I. However, there was no early indication of the beginning of the recoveries in 1970II and 1975II.

If we compare the first-quarter forecasts which predict the current quarter, the *ASA-NBER* survey has missed the direction of the growth in *GNP* in five quarters: 1970I; 1970II; 1970IV; 1974III; 1975II. It means that the *ASA-NBER* survey was unable to spot the turning points of 1970II, 1970IV, and 1975II, even though the forecast was made in the same quarter. In addition, the *ASA-NBER* first-quarter forecast made a false upturn signal in 1970I and 1974III when the economy did not really turn up. The Wharton premeeting first-quarter forecasts have made opposite forecasts of

the following four quarters: 1970II; 1974II; 1974III; 1975II. That means the first-quarter forecasts have missed the turning points of 1970II and 1975II, but gave false signal in 1974II and 1974III. The first-quarter prediction of the Wharton postmeeting forecast was more reliable in predicting the direction of the growth of real *GNP*. It has missed the turning points in 1970II and 1975II, but made only one false alarm in 1974II.

REFERENCES

V. Su and J. Su, "An Evaluation of ASA/NBER Business Outlook Survey Forecasts," in *Explorations in Economic Research*, Vol. 2, No. 4, Fall 1975.

A. H. Young, "Reliability of the Quarterly National Income and Product Accounts of the United States, 1947–71," Bur. Econ. Analysis staff paper, no. 23, Washington, July 1974.

On the Accuracy and Properties of Recent Macroeconomic Forecasts

By VICTOR ZARNOWITZ*

Most of the studies of economic forecasts cover very short time periods. Some attempt to grade forecasters on the evidence of how well they predicted change in a particular year or a few years. Yet it is clear that on any individual occasion some forecasters will be ahead of others by sheer chance or for some idiosyncratic reasons. The limitations of small sample studies of forecasts should be recognized. It is necessary to compile and examine forecast records extending as far back in time as possible, so as to gain information, take a longer view of forecasting behavior and performance, and place the evidence from short series of recent predictions in a proper perspective.

I. The Record of Annual *GNP* Forecasts Since 1947

In the early post-World War II period, most forecasts were made near the end of the calendar year for the next year and most referred to *GNP* in current dollars. The evidence we have on such forecasts goes back to 1947 but is quite fragmentary for the late 1940's and early 1950's.

The period of transition from the war economy witnessed the largest errors on record in the *GNP* forecasts. One reputable group of private forecasters came up with an average prediction for 1947 of a 6 percent *decline* in *GNP*, whereas the actual change turned out to be a *rise* of about 11 percent. For 1948 the group predicted a fractional decline but *GNP* instead advanced again at much the same surprisingly high rate. The failure of forecasts during these years was widespread; the developments of the time could not be predicted well with estimates based on data and relationships for the 1930's and false analogies with the early post-World War I period. When a recession finally came late in 1948, it proved shorter than many had expected, so the forecasters for 1950 erred again greatly on the pessimistic side.

The evidence for the period 1953–76 is summarized in Table 1 in terms of comparisons between the predicted and the actual annual percentage changes. The forecasts are those made late in the year $t - 1$ or, in a few cases, very early in the target year t. The actual changes used to compute the errors are based on the first official estimates for the year t published early in the following $(t + 1)$ year. These are provisional values which are themselves partly near-term predictions, and subsequent revisions are about one-third the size of the forecast errors. The errors are computed by subtracting the actual from the predicted changes, and they are as a rule negative (lines 5–8), which shows that the forecasts strongly tend to understate the changes (predominantly, increases) in *GNP* (on similar findings of earlier studies, see the author, 1972).

Table 1 discloses a substantial correspondence between the forecasts and the realizations. The predicted changes approximate the actual ones well in each period covered, the averages of the former being generally less than 1 percentage point smaller than the averages of the latter. The forecasts are in all cases considerably more accurate than a naive model which assumes that next year's percentage change will be the same as that of the previous year, and the somewhat less naive trend extrapolation model which projects the average percentage change of the four previous years (col. (4)).

The average error measures are important but fall far short of telling the whole story. Measures of correlation (which un-

*University of Chicago and National Bureau of Economic Research.

TABLE 1—SUMMARY MEASURES OF ERROR FOR ANNUAL PREDICTIONS OF PERCENTAGE CHANGE IN *GNP*, 1953–76

Line	Period and Number of Years Covered	Selected Private Forecasts, Mean[a] (1)	*Economic Report of the President* (2)	Wharton Model (3)	Extrapolation of Average Change[b] (4)
		Mean Absolute Error, in Percentage Points			
1	1953–76(24)	1.2			2.3
2	1956–63 (8)	1.5			1.9
3	1963–76(14)	0.9	0.9	1.3	1.8
4	1969–76 (8)	0.7	0.8	1.0	2.0
		Mean Error, in Percentage Points			
5	1953–76(24)	−0.7			−0.1
6	1956–63 (8)	−0.2			−0.4
7	1963–76(14)	−0.6	−0.2	−0.1	−0.5
8	1969–76 (8)	−0.4	0.2	−0.1	−0.5
		Squared Correlation (r^2) Between Predicted & Actual Change			
9	1953–76(24)	.79			.05
10	1956–63 (8)	.52			.04
11	1963–76(14)	.78	.75	.69	.08
12	1969–76 (8)	.86	.83	.75	.00

[a] Sources: Livingston survey, mean; *Fortune* magazine; Harris Trust and Savings Bank; IBM Economic Research Dept.; Nat. Securities and Research Corp.; Conference Board Economic Forum; R. W. Paterson, Univ. of Missouri; Prudential Insurance Co. of America; UCLA Business Forecasting Project; N.Y. Forecasters Club survey, mean (1956–63 only); *ASA/NBER* survey, median (1969–76 only).

[b] Assumes that next year's percentage change will be the same as the average percentage change in the four previous years.

fortunately are often omitted from forecast evaluations) are needed to show how well the predicted changes have tracked the actual changes over time. The r^2 coefficients for the forecasts covered in Table 1 are all positive and significant, generally exceeding 0.5 and, for the more recent periods, averaging 0.7 or higher (lines 9–12). In contrast, the corresponding coefficients for the extrapolations are generally close to zero.

The evidence supports the conclusion that the end of year forecasts of current-dollar *GNP* next year had a reasonably satisfactory record of accuracy since 1953. Indeed, in comparisons with earlier forecasts (see the author, 1967), that record improved considerably in the 1960's and even in the 1970's, a turbulent period presumed to have been particularly difficult to forecast.

More detailed inferences concerning the relative accuracy of the different forecast sets covered cannot be drawn from these results. One reason is that the forecasts differ appreciably with regard to their precise dates, and it is known from previous research that the earlier predictions have a significant advantage over the later ones (see the author, 1967; Stephen McNees, 1975). It is relevant, however, to make the general observation that the average error and correlation measures do not show large, consistent differences among the forecast sets being compared. This is in agreement with earlier findings, which strongly suggests that the search for a consistently superior forecaster is about as promising as the search for the philosophers' stone (see the author, 1971; McNees, 1975, 1976; Carl Christ).

To save space, some of the examined forecast sets are omitted from Table 1 and others are combined in the averages of

column (1), but the individual predictions for each year were analyzed in considerable detail. The year by year inspection revealed few major contrasts between the corresponding predictions, although the forecasters included certainly differ in many respects. Of course, competent forecasters use common data and techniques, regularly interact, and are often similarly influenced by recent events and current attitudes and ways of thinking. The genuine *ex ante* forecasts here considered are all to a large extent "judgmental," and this could well tend to reduce the dispersion among them; there is indeed some evidence that errors of *ex ante* forecasts with econometric models vary less than errors of *ex post* forecasts made without judgmental adjustments (see Christ). While published forecasts by ranking practitioners are often developed with particular skill or care, group average forecasts benefit over time greatly from cancellations of individual errors of opposite sign (see the author, 1967, 1972). At any given time, the deviations between corresponding forecasts from different sources are likely to be reduced by the working of these balancing factors. Thus, it is not surprising that forecasts for the same variable and target period tend to be similar. Indeed, the correlations between pairs of the forecast sets included in Table 1, computed for the four periods distinguished therein, are significantly higher than the correlations between predictions and realizations recorded on lines 9–12. The r^2 coefficients for the pairs of the predicted percentage change series all exceed 0.8, and some are considerably higher.

Of the individual observations comprised in the examined forecast sets, about 64 percent are underestimates and 34 percent are overestimates. By far most of the latter refer to years marked by economic recessions or slowdowns. The provisional *GNP* values show but two year-to-year declines in the period covered in Table 1: in 1954, which the forecasts overstated, and in 1958, which the forecasts missed (accounting for the only turning point errors in this sample). Thus underestimation was limited to the increases in *GNP*; moreover, it was most pronounced when the increases were particularly large.

Are these errors "systematic" in the sense of a bias that could have been readily escaped or corrected in advance? Not necessarily, although it seems difficult to discount them as merely another manifestation of the familiar tendency of forecasts to underestimate the observed changes (which, for series with random elements, is a property of even unbiased and efficient forecasts; see Jacob Mincer and the author, Michio Hatanaka). What is underestimated here is the average annual rate of growth in a series which, as properly recognized by the forecasters, is trend dominated and seldom declines from year to year. This outcome can be traced to the forecasters' tardy recognition of high-growth phases (booms) and, increasingly, of inflation speedups, but it was also mitigated by their even tardier recognition of business recessions and slowdowns. Such movements are recurrent and not purely random, and they have important, detectable regularities as shown by historical studies of business cycles; but they also vary a great deal over time, so their predictability remains very limited. In any event, simple "learning from past errors" would not have been of much use here as the errors of these forecasts generally have zero or very low autocorrelations.

II. Annual Forecasts of Real *GNP* and the Price Level

It is difficult to obtain and verify consistent forecasts of *GNP* in constant dollars and the implicit price deflator (*DIP*) that would cover more than just the most recent period. Few business forecasters in the 1950's and 1960's made systematic efforts to decompose their predictions of current-dollar *GNP* into quantity and price elements. Of the forecasters with econometric models who paid more attention to real *GNP*, only two (Michigan and Wharton) have longer records.

Table 2 shows that the forecasters predicted the real growth rates within

TABLE 2—SUMMARY MEASURES OF ERROR FOR ANNUAL PREDICTIONS OF PERCENTAGE CHANGES IN REAL *GNP* AND THE PRICE LEVEL, 1959–76

		Forecasts of Real *GNP*					Forecasts of the Implicit Price Deflator (*IPD*)				
Line	Period and Number of Years Covered	*Economic Report of the President*[a] (1)	Michigan Model (2)	Wharton Model (3)	*ASA/NBER* Survey, Median (4)	Extrapolation of Average Change[b] (5)	*Economic Report of the President* (6)	Michigan Model (7)	Wharton Model (8)	*ASA/NBER* Survey, Median (9)	Extrapolation of Last Change[c] (10)
						Mean Absolute Error, in Percentage Points					
1	1959–67 (9)		1.0			1.7		0.7			0.3
2	1962–76(15)	1.1	1.4			2.6	1.0	1.0			1.3
3	1969–76 (8)	1.2	1.6	0.9	1.0	3.6	1.4	1.4	1.4	1.3	2.0
						Mean Error, in Percentage Points					
4	1959–67 (9)		−0.5			−1.1		0			−0.1
5	1962–76(15)	0.6	0.2			0.1	−0.5	−0.5			−0.2
6	1969–76 (8)	0.8	0.8	0.5	0.7	0.7	−0.6	−0.9	−0.6	−0.9	−0.2
						Squared Correlation (r^2) Between Predicted and Actual Change					
7	1959–67 (9)		.53			.00		.42			.36
8	1962–76(15)	.78	.62			.05[d]	.77	.68			.54
9	1969–76 (8)	.86	.71	.94	.94	.32[d]	.58	.45	.60	.53	.17

[a] The forecasts for 1962, 1963, 1965, and 1968 must be inferred from statements in the *Report*, but they are confirmed by the Council as approximately correct (see Geoffrey Moore). The other forecasts are all based on figures given in the *Report* and so are fully verified.

[b] See fn. b, Table 1.

[c] Assumes that next year's percentage change will be the same as that of the previous year.

[d] r is negative.

average margins of about 1 to 1.5 percentage points (cols. (1)–(4)). The mean absolute errors of mechanical extrapolations of past percentage changes in real *GNP* are considerably larger (as illustrated in col. (5)). The mean errors of forecasts and extrapolations are negative for 1959–67 but positive in the more recent periods (lines 4–6). Correlations between the predicted and actual changes are all significantly positive, and they suggest some improvement in recent years: the r^2 coefficients for 1969–76 are higher than those for the earlier and longer periods (lines 7–9). Interestingly, they tend to exceed the corresponding coefficients for current-dollar *GNP* forecasts. In contrast to the reasonably high correlations for the forecasts proper, those for the extrapolations are here again extremely low or negative.

These summary measures, then, present the annual forecasts of real *GNP* in a generally favorable light. However, the accuracy of these forecasts varied greatly in different years, which at times impaired seriously their usefulness, and this does not show up in the summary. Real *GNP* turned down in 1954, 1958, 1970, and 1974, but eight of ten predictions for these years specified continued rises. Again, and not surprisingly, nearly all of the significantly large overestimation errors refer to the years during which national output grew at relatively low or decreasing rates, and most of the larger underestimation errors refer to the years of high real growth rates.

It is of considerable interest to note that the turning point errors are on the average as much as 2.5 to 3 times larger than the other errors in the real *GNP* forecasts. Hence, even though they occurred in only about 13 percent of the total *number* of forecasts, these errors had a strong adverse impact. Turning point errors account for 29

percent of the total absolute error of all real *GNP* forecasts in our collection, under- and overpredictions for 47 and 24 percent, respectively. This evidence contradicts the argument that such errors matter little because they are few and far between. It should be added that, relative to annual forecasting, in quarterly multiperiod forecasts turning points are more frequent and more difficult to predict and errors associated with them are more important. In this case missing a turn often means that a whole chain of predictions for the subsequent observations err badly. In sum, there are indeed strong reasons for makers and users of economic forecasts to give a great deal of attention to turning point errors. Actually, most of them realize this, as shown by the widespread practice of analyzing such errors (see Bert Hickman; studies cited in Gary Fromm and Lawrence Klein). However, there is certainly much need for improvement here, and room for some new initiatives (for example, on how to use current signals from leading indicators see Beatrice Vaccara and the author).

The worst single year for the predictions covered in Table 2 was 1974, on the eve of which forecasters across the field missed the onset of a serious recession. This, plus the smaller turning point errors for 1970, are the main reasons for the rise in the average errors of these forecasts in 1969–76 compared with the earlier years. But the rise in the absolute errors was not large, and there was no decline in accuracy as measured by the criteria of comparisons with extrapolations and correlations of predicted with actual changes (see Table 2, cols. (1)–(5)).

Table 2, cols. (6)–(10), surveys the performance of forecasts of percentage changes in *IPD* that match the real *GNP* predictions covered in cols. (1)–(5). Here again, the mean absolute errors are close to or somewhat higher than 1 percentage point (lines 1–3). On the average, the predicted inflation rates fall short of the actual ones by fractions of 1 percentage point (lines 4–6). The 1959–67 forecast sets are less accurate than simple last-change extrapolations (col. (10)), and the other sets outperform the naive models by relatively small margins, much less than those observed for the *GNP* series. (For *IPD*, unlike for *GNP* in current and constant dollars, projections of the last change have smaller errors than those of the average change and hence are used here.) The forecasts underestimated strongly, much more so than the extrapolations, the average inflation since 1961. The predicted and actual percentage changes in the price level are all positively correlated, but the correlations for 1969–76 are generally lower than their counterparts for *GNP* and, still more so, for real *GNP*.

Forecasts of inflation often have much in common with projections of the last observed rate of inflation. For the four sets of forecasts in Table 2, cols. (6)–(9), the r^2 coefficients of the correlations between their errors and the errors of the corresponding extrapolations range from 0.51 to 0.95 and average 0.76. Hence, like the last-change extrapolations, the inflation forecasts tend to lag a year behind the actual rates of inflation. Indeed, the correlations between the predicted changes and the previous year's actual changes are all positive and high: the r^2 coefficients for our four sets of *IPD* forecasts vary between 0.72 and 0.87 and average 0.79.

A fair inference from these results is that the forecasts of inflation are indeed poor, a finding which is also consistent with other evidence. Improvements will require major advances in our knowledge, presumably through research that would be solidly based on carefully worked out data, since abstract speculation abounds but good information and observation are rare in this area.

The annual percentage changes in real *GNP* are inversely related to those in *IPD* and positively related to those in current-dollar *GNP*, while the last two variables do not show a strong or stable association. The relationships between the predicted changes generally parallel the actual ones. The errors of the forecasts are similarly interrelated, as shown by the tabulation in Table 3 of r^2 coefficients (*RGNP* denotes real *GNP*; the corresponding r coefficients

TABLE 3—SQUARED CORRELATION (r^2) BETWEEN FORECAST ERRORS

Source of Forecast	for *RGNP* and *IPD* (1)	for *RGNP* and *GNP* (2)	for *IPD* and *GNP* (3)
1962-76 (15 years)			
Economic Report (*CEA*)	.297(−)	.359	.114
Michigan model	.494(−)	.429	.006
1969-76 (8 years)			
Economic Report (*CEA*)	.677(−)	.004	.259
Michigan model	.684(−)	.209	.014
Wharton model	.340(−)	.036	.466
ASA/NBER survey, median	.524(−)	.013	.351

are positive, except where the sign (−) indicates the contrary). The pervasive pattern of negative correlation between errors in forecasting real growth and inflation (col. (1)) deserves an emphasis.

III. Concluding Observations

The end of year forecasts of annual percentage changes in *GNP* earn good marks for overall accuracy. Moreover, they are found to have improved in the period since the early 1960's compared with the previous years after World War II. The real growth (*RGNP*) and inflation (*IPD*) forecasts are less accurate. The former suffer from large turning point errors, the latter from large underestimation errors. The errors in predicting real growth are negatively correlated with the errors in predicting inflation, which helped to make the nominal *GNP* forecasts more accurate. In recent times, these correlations were connected with the unexpected concurrence of accelerating inflation and slowing, then declining output rates: optimistically, and probably also from a lingering faith in a simple Phillips tradeoff, forecasters kept anticipating less inflation and more growth. But in the late 1950's and early 1960's, it was the relative stability of the price level that caused widespread surprises and offsetting errors resulted from the opposite combination of overestimates of inflation and underestimates of real growth.

The favorable record of annual *GNP* predictions does not imply that forecasters can perform well the more difficult task of predicting quarterly changes in *GNP* within the year ahead or even beyond it. An examination of the recent multiperiod predictions shows that the errors for real *GNP* and *IPD* cumulated rapidly beyond the spans of two to four quarters. Previous studies have shown the cumulation to be as a rule less than proportional to the increase in the span, but in the 1970's the forecast errors build up much faster than usual. At least in such turbulent times, the predictive value of specific numerical forecasts reaching out further than a few quarters ahead must be rather heavily discounted.

REFERENCES

C. F. Christ, "Judging the Performance of Econometric Models of the U.S. Economy," *Int. Econ. Rev.*, Feb. 1975, *16*, 54–74.

G. Fromm and L. R. Klein, "The NBER/NSF Model Comparison Seminar: An Analysis of Results," *Annals Econ. Soc. Measure.*, Winter 1976, *5*, 4–5.

M. Hatanaka, "The Underestimation of Variations in the Forecast Series: A Note," *Int. Econ. Rev.*, Feb. 1975, *16*, 151–60.

Bert G. Hickman, *Econometric Models of Cyclical Behavior*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 36, New York 1972.

S. K. McNees, "An Evaluation of Economic Forecasts," *New England Rev.*, Nov./Dec. 1975.

———, "An Evaluation of Economic Forecasts: Extension and Update," *New England Rev.*, Sept./Oct. 1976.

J. Mincer and V. Zarnowitz, "The Evaluation of Economic Forecasts," in Jacob Mincer, ed., *Economic Forecasts and Expectations*, New York 1969, pp. 14–20.

G. H. Moore, "The President's Economic Report: A Forecasting Record," *Nat. Bur. Reporter*, New York, Apr. 1977.

B. N. Vaccara and V. Zarnowitz, "How Good Are the Leading Indicators?," 1977 *Proc. Amer. Statist. Assn., Bus. Econ. Statist. Sec.*, Washington forthcoming.

V. Zarnowitz, *An Appraisal of Short-Term Economic Forecasts*, Occas. Paper 104, Nat. Bur. Econ. Res., New York 1967.

———, "New Plans and Results of Research in Economic Forecasting," in *New Directions in Economic Research*, Nat. Bur. Econ. Res. *51st Annual Report*, New York 1971.

———, "Forecasting Economic Conditions: The Record and the Prospect," in his *The Business Cycle Today*, New York 1972.

DISCUSSION

OTTO ECKSTEIN, Data Resources, Inc. AND PAUL M. WARBURG, Harvard University: The three papers provide some comfort to the forecaster. They show that short-term macro forecasting has produced results that are substantially better than extrapolation methods. It would be a sad day if this were not the case, since forecasting is one test of whether the body of macroeconomics, embodying the human capital of so many capable thinkers and researchers, has a positive social product.

The three papers are done with high craftsmanship and objectivity. The auditing of forecasts help improve their accuracy by maintaining the "best effort" of the forecasters, and by producing technological progress. We hope that the work of Stephen McNees, to whom we all have to be grateful for developing the track records for the 1970's, will be maintained, and we are glad to welcome back Victor Zarnowitz for a second major round of work in his forecast evaluations.

To raise just a few specifics: Vincent Su contrasts the results of the *ASA-NBER* forecasters survey with the forecasts produced by the Wharton group, and uses these two sources to compare econometric with noneconometric forecasts. Perhaps in the 1950's and early 1960's there was validity to this distinction, but it is long gone. We doubt that there is any member of the *ASA-NBER* panel who is not affected by the forecasts of the econometric model groups, and conversely, econometric model forecasting has assimilated the informal forecaster's grist of surveys, data, and policy analyses. It is not surprising that econometric models have become central to virtually all serious forecasting. It would be strange indeed if computer technology—the central technological change of our era—did not have a massive impact on economic forecasting which is essentially an information processing activity.

We disagree with Zarnowitz on one critical point. He abjures adjustments of forecasts for base revisions, assuming that revised data represent the "truth" and that all differences between preliminary and revised data are reductions in measurement error. This is a questionable assumption. Data revisions have at least three principal sources: more complete data become available; benchmark data, usually of cross-section survey origin, replace time-series interpolators; and government agencies redefine concepts. The practical forecaster soon discovers that much of the government's revision work is counterproductive for forecasting. Completeness of coverage of later data principally means the inclusion of smaller economic units through sampling programs, whereas preliminary data mainly reflect the larger economic units with better reporting systems. Benchmark cross-section data may indeed improve long-term measurement, but at the expense of time-series consistency. Taken at infrequent intervals, with changed procedures and personnel both at the issuing and at the form-filling end, they cannot be consistent from one survey to the next. Yet forecasts, whether by econometric model or informal method, rest almost entirely on the time-series properties of the recorded data. Further, the government's conceptual "improvements" of recent years have come at the expense of the data's information content. For example, this month's third-quarter *GNP* revisions, which originated in the reclassification of some stores selling to farmers from the retail to the wholesale category, may have a sound conceptual basis but are surely not helpful in assessing the 1977 economic developments as a basis for forecasting. The changing treatment of profits and depreciation in the official data are larger examples of revisions that reduce information content.

The paper by McNees finds that the extent of bias in forecasts is rather small. While we are relieved that our forecasts are not "irrational" according to the canons of the rational expectations theorists, we do not take much comfort in this finding. Our forecasts are not sufficiently cyclical, systematically understating the violence of

change over the business cycle. It is only the poverty of this definition of rational expectations which lets us pass the test; a definition which would reflect the objective functions of forecast users would include not just biases in the levels of forecasts, but also in the higher moments of the forecast path's distribution.

Let us now summarize some findings of our own on the current status of forecasting: The forecasts of 1974 were the most serious forecast error since the false perspectives found near the end of World War II. The forecasts failed to anticipate the violence of the decline, September 1974 to January 1975, which was both the worst financial crisis and the most dramatic downward revision of near-term sales and profit expectations since the Great Depression. The financial crisis was made possibly by the collapse of the Nixon presidency which destroyed the usual check and balance on monetary policy. The collapse of expectations was due to the belated perception that some of the investments of the previous boom had been irrational, and that the new world oil situation had rewritten the rules of the game of economic growth. Forecasting methods that existed at that time were inadequate to anticipate these matters. In response to this forecasting failure, DRI rebuilt its *U.S.* model to incorporate more of the financial and raw material factors that helped to create that collapse.

Since that collapse, the record of forecasting is far better than it was in earlier periods. The models showed very early that the collapse would be followed by an upturn that would be vigorous by the consumer and moderate by business, and that the economy would reliquify, relieving financial strains and permitting high housing activity.

The principal problem of short- and intermediate-term forecasting continues to be the projection of the upper business cycle turning point. The difficulty lies in the structure of the economy. The upper turning point in every postwar business cycle has followed the sudden development of a credit crunch. Before the crunches, there was no reason to look for the turning point. But the interval between the diagnosis of the crunch and the upper turning point is very short.

The leading indicator approach of the *NBER* and the U.S. Bureau of the Census has been the principal analytical device for identifying business cycle turning points. But the accomplishments of this method are also fairly slim. For example, leading indicators began to fall in August 1974—information became available in late September. The three months of data necessary to confirm a decline were not available until late November.

The principal challenge to short-term forecasting is to improve the methods for predicting upper turning points. A combination of monitoring systems and models holds out the best hope, we believe.

The Business of Business is Serving Markets

By JOSEPH L. BOWER*

When I first studied economics in 1955, it was taken for granted that the subject matter of the field was the allocation of resources and the distribution of income. The beauty of the field derived from its relative elegance as a social science. A limited number of assumptions concerning the behavior of firms and the behavior of individual consumers permitted the construction of an elaborate set of hypotheses concerning markets and the economy. Particularly enchanting for me was the application of micro theory to investment decisions in the firm.

From the perspective of 1977, I believe that the reality of firms and markets has changed sufficiently to raise extensive questions as to the usefulness of our traditional ways of using economic models for the analysis of firms, industries, and markets. Specifically I believe that Alfred Chandler's *The Visible Hand* is often replacing the invisible hand as the most efficient allocator of resources; second, a domestic view of most markets is inadequate from the point of view of political economy; third, the notion of product as traditionally conceived is inadequate: consumers are buying far larger bundles of attributes than normally considered, with substantial consequences for the notion of competition and the careful definition of industry boundaries. In addition, much of what firms sell today are public goods with substantial consequences for how we view both the firm and the market.

I. The Visible Hand

In his book, Chandler described the development of industry in the United States as a process by which markets were internalized by growing firms seeking efficiency. The process of vertical integration that he described has been followed by a pattern of diversification that has equally important consequences for the economy. The large modern firm that accounts for the vast bulk of the production of our *GNP* is a product market diversified organization.

General Electric (GE) is perhaps the prototype of a widely diversified firm. It produces an enormous range of products in both the industrial and the consumer sectors of the economy. Despite the diversity of the firm, its management seeks to review the substance of individual subunit strategy. Great attention is paid to the potential interrelationship among some businesses. Constant effort is made to be sure that human resources are allocated optimally. The business in trouble will soon find itself under the management of a GE trained executive with skills his superiors feel are ideally suited to the needs of that troubled unit's situation. Large functional staffs at headquarters back up the efforts of group and corporate executives to discipline and critique the operating and strategic plans of the nearly one hundred $100 million businesses GE operates.

Another very different variety of diversified firm is exemplified by Textron. At Textron, explicit attempts are made to make sure that there are *no* interrelationships other than financial between the subunits of the business. Textron can thus be considered a forward-integrated investment bank, or perhaps a backward-integrated consulting firm. The corporate staff in Providence numbers under 200 including secretaries. Obviously, the corporate office of Textron has a very different concept of how it is managing than does GE.

*Graduate School of Business Administration, Harvard University.

The differences between GE and Textron are typical of the remarkably distinct approach to management adopted by what Richard Rumelt called "related diversified firms" on the one hand, and "unrelated diversified firms" on the other. These differences are exemplified by Norman Berg's study of strategic planning and conglomerate companies. The absence of any functional staff other than finance and control in the unrelated diversified company reflects a totally different approach to corporate intervention in divisional affairs. When a subunit is in trouble a corporate management can only try to find new general management or sell the division. Its concept of management does not include the application of corporate level substantive expertise.

The problem of managing such extreme diversity has led the analysts at firms such as GE, as well as their consultants, to develop a number of concepts designed to make subunits comparable for purpose of resource allocation. The application of these concepts (independent of whether they are tested or valid) is a strong example of the internalization of the market that I am describing here.

Three particularly important concepts of current interest and use are the product life cycle, the experience curve, and the portfolio approach to the problems posed by a maximum sustainable rate of growth. Let us consider each briefly.

A. *The Product Life Cycle*

A considerable body of research has discovered that the market life of a product is typically characterized by an early phase in which the growth of sales is slow while the market learns about the product; a second phase in which growth is extremely rapid corresponding to the most rapidly ascending portion of an "S" curve; a third phase in which growth slows; and a fourth phase in which sales may actually decline. Appropriate competitive behavior during these phases has been studied quite carefully and there is substantial evidence that patterns of successful response do exist. On the other hand, it is also clear that the efforts of firms can change the shape of the product life cycle with respect to time. Careful and aggressive work can extend the growth phase of the curve. For example, over a period of thirty years, Dupont has continued to find new uses for nylon monomer. It thereby has kept the growth of nylon on a very steep slope. In contrast, in some industries the introduction of substitutes is so rapid that the life cycle is relatively short. The semiconductor firms seem to be following this practice.

B. *The Experience Curve*

A second and related concept is the notion of the experience curve. Studies of various industries conducted by the Boston Consulting Group revealed that total average unit cost declined logarithmically with cumulative output. In fields as varied as germanium transistors, integrated circuits, motor gasoline, ethylene, benzene, propylene, primary aluminium, primary magnesium, black and white television receivers, facial tissue, electric power, and motorcycles, it has been shown that unit costs decline in constant proportion to cumulative volume.

The source of this phenomena lies beyond traditional economies of scale. It relates in fact to a phenomena that is discussed below, that is, the extent to which a consumer actually buys a considerably larger bundle of attributes than is naturally associated with the product. The experience curve shows the unit costs declining because a large number of activities accounted for as overhead, but productive from the point of view of a customer, can be spread over greater volumes. At the same time, there is a learning effect, normally seen only in a factory, whereby the same product can be produced with equal or superior quality, simply because the process of making that product is better understood.

Taken together with the product life cycle, the experience curve has an awesome implication. At that point in a product's life where its market begins to

take off, the firm which seeks to have the lowest cost position must invest as if there were no tomorrow. It becomes very important to capture a substantial market share, so that one's cumulative volume grows faster than that of any other competitor. The alternative is a perpetually higher cost position.

C. *The Business Portfolio*

The problem for the diversified firm can now be posed quite sharply. Often a company is lucky enough to be engaged in two or three such rapidly growing businesses. Where can it find the funds? In fact, its growth is limited to a maximum rate easily expressed in terms of the corporate return on investment, the borrowing policy, and the dividend policy for the firm. Within that constraint, we can imagine the portfolio of opportunities facing the firm with a simple 2 by 2 matrix.

The Boston Consulting Group's presentation of this problem uses market share as a proxy for relative cumulative experience. That is, businesses with high relative market share are assumed to have comparatively high cumulative experience.

		Relative Market Share	
		High	Low
Market	High:	*	?
Growth	Low:	$	Dog

Then, it is generally true that high market share, high-growth businesses are outstanding producers of earnings, although on balance they may absorb more cash than they generate (they are stars). High-share, low-growth business should be profitable and generate cash (they are cash cows). Low-share, low-growth businesses are often very dismal propositions, sometimes incapable of correction (they are dogs). Low-share, high-growth businesses are a different story. The risks to which investments are subject are high, but if the firm's skills can be used to capture a relative high share, then profits are proportionately great. An example of these wildcats is Memorex which joined the *Fortune* 500 after only 15 years of existence by plunging heavily in computer tape, disk packs, and disk drives. Catching the products early enough in their life cycle, Memorex was able to establish a competitive cost position with 3M and IBM.

My first point can now be made. The *U.S.* economy today is populated by a majority of large firms that are regularly scanning their portfolios of businesses trying as best as they know how to anticipate market growth; competitors' behavior; changes in technology; and whatever other factors will influence 1) their chance to convert a question mark to a star, 2) their ability to rescue a dog and, 3) the appropriate moment to start harvesting the cash from a star becoming a cash cow. It is a far more disciplined resource allocation process than the capital markets provide.

I believe it is no accident that Youngstown was shut down by a conglomorate rather than a steel company. The steel company's alternate uses of capital were different from and less attractive than those facing Youngtown's conglomerate parent.

II. International Markets

One of the most dramatic vignettes in the current world economic scene is the spectacle of Zenith suing the *U.S.* government to enforce dumping laws on TV sets, and the Consumers' Union fighting to protect imports. What is happening?

In color TV, the Japanese used their position in their large home market to drive down the experience curve. Net wages are nearly the same as for the *U.S.* industry. Yet, by applying superb process engineering they have achieved both lower costs and higher reliability than their *U.S.* competitors. Sony, Matsushita, Hitachi, and Toshiba enjoy roughly a 75 percent world market share. When their higher yields permitted introduction of a one-year warranty (*U.S.* practice was 90 days), the warranty cost for one prominent *U.S.* firm allegedly rose from 2 to 9 percent of sales.

Motorcycles provide a more dramatic example, perhaps because we have more data. A 1975 study produced for the British House of Commons by the Boston Consulting Group revealed the impact of Japanese economics on a traditionally strong British

industry—motorcycles. In 1960 the Japanese produced large numbers of small bikes but exported only 4 percent of their production. They used this base, however, to invade the low end of the British market. Less sophisticated in both manufacturing and marketing, the British withdrew from the low-price segment. The Japanese developed volume and experience in progressively large models, displacing the British who sought short-term profit by withdrawing from these price competitive markets. Followed throughout the world, the consequences of this strategy are market shares of 87, 88, 74, 74, and 70 percent in the United States, Canada, United Kingdom, France, and Germany.

The volume base this gives Japanese production is such that the British study estimated all but the smallest of Honda's products in the United States to be selling at a premium of 24–43 percent over prices in Japanese markets. The return on investment for the Japanese producers is high.

The same Japanese strategy has succeeded in jewelled watches, cameras, automobiles, as well as steel. It has not succeeded in pin-lever watches, in chemicals, in drugs, in semiconductors, or in major home appliances, to take a wide range of examples. Why? Has fragmented competition in the United States produced efficiency? No. In pin-lever watches, Timex simplified the design, then mass produced and marketed a quality product. They crippled a Swiss industry responsible for 8 percent of Swiss employment. In chemicals, the major *U.S.* manufacturers practiced experience curve strategies for years. They are the leading producers in the world. In drugs, Japanese industrial engineering—key to low cost in assembly based manufacturing—has left their drug companies totally dependent of *U.S.* and European patents. In semiconductors, Texas Instruments and a few *U.S.* competitors have practiced experience curve pricing with a vengeance even the Japanese respect; *U.S. R & D* is also formidable.

In major home appliances, GE, Whirlpool, and Design and Manufacturing have made the investments in efficient manufacturing process and scale to block any Japanese advantage. In fact, TV and major appliances are a dramatic and instructive contrast. In one industry imports have captured virtually the entire *U.S.* market. In another—similar with respect to manufacturing and distribution—the *U.S.* manufacturers are preeminent. The consequences of superior strategy for *U.S.* balance of payments and employment are obvious.

The point, however, should be clear. No market ought to be considered in solely domestic terms until study justifies that conclusion. Schumpeter's gale of competition blows from all points on the compass but particularly from Europe and Japan. In Japan, it is primarily aggressive, modern, volume, and experience conscious companies exploiting *all* the world's markets that account for that country's success—*not* Japan, Inc.

III. The Concept of Product—There's More to it than What You See

The final problem with the traditional economic view is that it badly oversimplifies the concept of product, especially in empirical analysis. There are two major dimensions to the problem—private and public.

A. *What the Consumer Buys Defines the Product*

Earlier mention was made of Timex. By 1955 Swiss standards a Timex wasn't a watch. Swiss firms were selling jewelled precision timekeeping instruments. Later, after Timex took half the market, they discovered that most customers were buying adequate reliable timekeeping while a few were buying jewelry.

In the same spirit as the Swiss, economists insist on leaving Sears, Roebuck and Company out of the major home appliance industry. The fact is that the customer buys features, reliability, and service at a price. In the range of functions from design to manufacturing to distribution to selling to service, Sears performs all but the manufacturing function. In contrast, GE performs all but the retail selling function

(they do sell to home builders). A small company such as Tappan neither retails nor services. As Michael Hunt has shown, the strategies of GE *and* Sears shape the industry. Sears has perhaps the greatest impact of all. It *is* in the industry, *if* that industry is properly defined.

Another example of the same phenomenon is the so-called ego-intensive good, such as cosmetics and fashion garments. Here, the retailer is a key ingredient of what the consumer buys. Seeking to enhance their own self-image, the customer needs to be reassured that the product is "right." The increase in price of a T-shirt with "Bloomingdale" on it is a simple example of the point.

Still another example is the clear preference of airline customers for the airline on a route with the most flights scheduled. Market share increases disproportionately with capacity share. In industrial goods excess capacity can have a similar impact. Customers will buy speed of response as well as product.

Perhaps the most dramatic example occurred in computers. There, in the early years of the industry, a number of firms made awkward attempts to sell machines to customers. What was true then, and seems just as true today, is that most customers want to buy reliable electronic data processing—they don't want machines. The customer's concept of the market defines the market *not* the producer.

The problem for traditional economics is that the definition of "producer" gets complicated. What are we to make of manufacturers without factories? How do we measure concentration? What does it mean that the vertically integrated manufacturer with assured source of supply is selling something different than the single stage producer?

B. *The Product the Consumer Buys Often Includes Public Goods*

Still more difficult for traditional economics is the contemporary approach of Congress and state legislatures to the purchase of public goods. Whenever there is a choice between persuading the public to pay for a program directly through taxes and imposing the costs of the program indirectly through regulation of industry, our legislators choose the hidden approach.

Examples of this phenomenon are so extensive as to defy a complete listing, but a few illustrations are useful: 1) we are buying clean air and water with pollution laws; 2) we are buying racial integration of the society through equal opportunity laws; 3) we are buying political fragmentation through the antitrust laws; 4) we are buying redistribution of income through the exemption of unions from the antitrust laws; 5) we are buying better health through auto safety specifications.

Perhaps the most easy to understand and diagnose is the last example. Studies have shown that the leading source of auto accidents is the driver. He drives too fast and is often poorly trained. The next source is roads—poorly designed, poorly lit, and poorly signed—they cause problems. The next is the car itself. However, our auto safety laws attack car manufacturers, not the states, nor the drivers. While they have helped reduce death and serious accidents, their effects were dwarfed by the reduction in speed to fifty-five mph.

The point here is not that Congress may buy public goods in a way that is deceptive or not cost effective. The point is that we *are* now asking companies to produce a wide range of goods including—perhaps most important of all—stimulating, fulfilling, steady, high-paying, healthy jobs located within the fifty states.

Together, the three points set forth above sum up to the following picture. Our economy is dominated by large diversified firms seeking higher return by allocating resources to achieve a balanced portfolio of businesses. Each of these businesses is typically competing in a multinational market, often against giant European and Japanese firms. Finally, the products of these businesses are multidimensional including a wide range of public goods that are bought indirectly through regulation of private firms.

IV. Implications for Policy

The difficulty with the traditional view of the firm is nowhere more apparent than in our antitrust laws. At this moment in history we find: 1) virtually all governments in the world seeking the concentration of industry in pursuit of efficiency and defense of jobs; 2) our own economy under pressure to reduce pollution and conserve energy; 3) a major attempt underway to improve the nature of work and to broaden access to all jobs. At exactly this time, we are attacking precisely those firms that best meet our national objectives as efficient allocators of resources and/or effective providers of public goods.

The antitrust cases against IBM and Xerox seem to the point. Both firms are widely regarded as leaders in their products, management, and employment practices. They are powerful forces in their world markets; but nonetheless well regarded citizens of host countries. The Department of Justice case against IBM is built on an astonishingly narrow definition of the product and hence the competition. The case against Xerox attacks the success of its marketing practices. It seems clear that the Department of Justice would be happier if Fugitsu had some of IBM's market share, and the Federal Trade Commission pleased with Rank Fuji taking some of Xerox's share.

My purpose here is not to defend IBM or Xerox. Rather, I am concerned that an outdated vision of how a market works—or how political decentralization is preserved—can serve as the basis for attack on our best producers of public goods, our strongest international agents, and our best producers of private products. The *Visible Hand* is—and always was—a cost effective way of serving an important set of markets for economic and social goods. As I see the task for future researchers it is to learn the content and boundary of that set, to learn how to intervene where it is necessary, and to invent other mechanisms for producing those goods and services for which the *Visible Hand* is poorly suited.

REFERENCES

N. Berg, "Corporate Role in Diversified Companies," Grad. School Bus., work. paper HBS 71-2, Harvard Univ. 1971.

Alfred Chandler, Jr., *The Visible Hand*, Cambridge, Mass. 1977.

M. Hunt, "Competition in the Major Home Appliance Industry," unpublished doctoral dissertation, Harvard Univ. 1972.

R. Rumelt, "Strategy Structure and Economic Performance," Div. Res., Grad. School Bus., Harvard Univ. 1974.

Boston Consulting Group Ltd., *Strategies for the British Motor Cycle Industry*, London 1975.

On the Basic Proposition of X-Efficiency Theory

By HARVEY LEIBENSTEIN*

The view behind this paper is that although neoclassical (*NC*) micro theory works some of the time, there are areas of experience to which it is not applicable. As a consequence it is desirable to develop models which are more general than the *NC* framework, which fit economic realities, and into which the *NC* framework fits as a special case. In what follows I shall focus our attention on the intrafirm organizational elements of X-efficiency theory, and on what will be referred to as the basic proposition of that theory.

I

Suppose a multiperson firm is given the following option: to produce X-units for which it is offered successively larger budgets $B_0, B_1, \ldots, B_n$ ($B_0 < B_1, < \ldots < B_n$) plus a *fixed profit*. The firm is free to return any portion of the budget. What size budget will it choose? What happens to cost per unit as B increases? I believe that the best answer is that the firm would probably choose B_n, and cost per unit will increase in proportion to the increase in B. There is no benefit to returning any of the budget. Keeping it gives firm members more elbow room, since it would allow them to choose to work less hard or harder, give less attention to details or more attention, choose their own good time, rather than to feel pressured by time, etc. Certainly there is much casual empirical knowledge that fits this picture. We would not expect such a firm to minimize cost per unit; that is, to choose the lowest possible budget. This suggests that under parallel circumstances firms will not minimize costs for a given output unless competition or environmental elements force them to do so. By the degree of X-inefficiency I shall mean the excess of actual over minimum cost for a given output.

II. Noncost Minimization

The basic postulates and related variables of X-efficiency theory are indicated in Table I and contrasted with their neoclassical counterparts.[1]

A maximizer would take advantage of all opportunities for *gain*, and attend fully to *all* constraints which, if not attended to, would impose a *loss*. I shall use the term "constraint concern" for the degree of attentiveness to 1) opportunities for gain *and* 2) constraints which can impose losses. According to the selective rationality postulate, an individual's choice of the degree of constraint concern depends (a) on his personality and (b) on the economic context. Individuals are assumed to compromise between the way they feel they *ought* to behave (superego or standard of behavior) and the way they would like to behave (id, or unconstrained desires). Up to a point individuals will trade off less constraint concern for more internal *felt* pressure. However, economic contexts may contain more external pressure than one likes. With increases in *contextual pressure* we would expect increases in constraint concern. Thus contextual pressure and personality (taste for constraint concern) determine the extent of the deviation from maximizing behavior; that is, the degree of constraint concern.

The decision unit is the individual. He has to exert effort, which is made up of the following components: the activities A, the pace of the activities P, their quality Q, and the time sequence T. Thus individuals may be said to choose an effort point; that is, an

[1]The ideas in this section have been developed more fully in my book, but to enable this paper to be more or less self-contained, it seems desirable to sketch these notions at this point.

*Harvard University.

TABLE 1

Components	X-Efficiency Theory	Neoclassical Theory
1. Psychology	Selective rationality	Maximization or Minimization
2. Contracts	Incomplete	Complete
3. Effort	Discretionary variable	Assumed given
4. Units	Individuals	Households and firms
5. Inert areas	Important variable	None
6. Agent-principal	Differential	Identity of interests

APQT bundle. In general, a job interpretation will reflect a number of effort points, to be referred to as an effort *position*, which allows the individual to meet some changing demands on effort.

A component not contained in the standard neoclassical system is that of inert areas. By this I have in mind that individuals who find themselves in a given position will not necessarily move to a superior position in the standard utility sense because of the inertial cost of moving. The inertial cost depends on an individual's personality. Thus a maximizing individual would have a zero inertial cost. Of course, in specific contexts there may also be additional costs of moving from one position to another.

The basic argument as to why we should not expect cost-minimizing behavior in the multiperson firm (outside of *perfect* competition) is as follows. Because firm membership relations involve incomplete contracts in which the details as to what an individual does for the firm are not completely specified, effort discretion exists. To interpret his job a firm member exercises effort discretionary options. He chooses an effort position. Because of effort discretion the firm member can choose noncost-minimizing positions. Also, because of differential principal-agent interests we would normally expect most individuals to choose noncost-minimizing positions. Furthermore, the existence of inert areas will imply some degree of persistence of such positions. Finally, since we do not assume maximizing behavior, there is no reason for anyone in a firm made up of agents to try to impose cost-minimization effort points on themselves or others.

III. Organizational Entropy—The Rising Cost Tendency

A central facet of my model is organizational entropy, which implies a cost-rising tendency that management has to struggle against. The essence of the entropy phenomena is that effort decisions become less and less directed toward presumed management objectives and costs rise as a consequence. We can see the nature of the entropy phenomena if we keep in mind that the original effort decision is made in the light of beliefs about contextual constraints; that is, in the light of the controls and influences of others on the quality and nature of the individual's effort. Interpersonal relations within the firm involve implicit monitoring. But as others in the firm reduce their controls and influences, the individual in question alters his beliefs about contraints and reconsiders his effort position.

A way of characterizing organizational entropy follows: Suppose the effort position E is decomposed into a sum of simpler effort elements $e_1, e_2, \ldots, e_n$. Compare this with E_{n-1} which contains only the elements $e_1, e_2, \ldots, e_{n-1}$ in such a way that E_{n-1} is a less effective effort point than E_n. The effort elements deleted reflect such phenomena as being less careful in carrying out activities, dropping activities which would improve quality, eliminating activities which increase pace, etc. Hence, shifts in effort positions from E_n to E_{n-1}, $E_{n-2}, \ldots$ define the process of effort entropy. Of course, parts of the activities help to control and influence (for example, monitor) the activity of other individuals. Thus, the reduction of effort by one indi-

vidual involves reducing the constraints faced by others, which in turn will have the tendency to reduce the value of effort by others.

Organizational entropy can also be seen in terms of the *effort responsibility consequence* (*ERC*) relations between those who make effort choices and those on whom they have an impact. Consider two extreme cases. For the one-man firm (the complete responsibility case) the entrepreneur, manager, and worker are all the same individual. All trades take place in the market. Every individual is completely responsible for the consequences of his effort. If his effort is less, he has less of value to offer, and he receives less in return. There are no losses (or costs) that are imposed on others. This is the neoclassical micro-theory case.

Now consider the other extreme, complete irresponsibility. All firm members are hired on a time basis; they are not responsible to the "stockholders." There is considerable effort discretion. If an individual were asked whether he would like rules under which he is responsible for the consequences of his activities, or rules under which he is not responsible, he would choose the latter. As a result, activities (or parts of effort positions) which impose costs on the organization are passed on to others and ultimately either to stockholders, consumers, and/or government, where governments provide subsidies. Another way of looking at the matter is to note that in the case of effort discretion, with wages predetermined, free-rider incentives exist.

IV. The Basic Proposition of X-Efficiency Theory

Using the ideas developed in the previous paragraphs I develop what may be viewed as the basic proposition of X-efficiency theory. Individuals' contracts and effort choices impose costs on the firm, and the firm attempts to obtain "budgets" from the "environment" to at least cover such costs. It is assumed that individuals are so motivated that they prefer less confining constraints to more confining ones. This translates to a desire for greater rather than smaller individual budgets.

Suppose that the firm can be given one budget out of a set of budgets $B_0 < B_1 < \ldots < B_n$, to cover the cost of producing quantity X. The minimum budget B_0 is estimated on the basis of the cost of production of X units under perfect competition. The budget B_i can be allocated among firm members: b_{ij} is that allocation of budget B_i to firm member j, and $\Sigma_j b_{ij} = B_i$.

Let us now examine the intrafirm situation. Local effort discretion is covered by the *ERC* relations for individuals;[2] i.e., ERC_j for all j. The context allows individuals to use capital and other services, which added to the wage determine the direct cost d_{ij} imposed by individual j. Also note that individuals can also impose costs on others (indirect costs i_{ij}) by negatively influencing the effectiveness of the effort of others.

An *ERC* relation is said to be directly *tight* if the individual cannot avoid any of the direct cost consequences of his effort choices; that is, if others in the firm determine his wages or control his effort so that he is always completely financially responsible for the results of his effort. For example, this is the case when the wage is always equal to the value of his marginal product. The opposite of tightness is looseness. In addition, the effort may also increase the costs that result from the impact of the individual's effort on the efforts of others. By definition the costs imposed by an individual $c_{ij} = d_{ij} + i_{ij}$. Also c_{ij} can be $\geq b_{0j}$. Thus direct tightness implies $d_{ij} = b_{0j}$. Indirect tightness implies $i_{ij} = 0$.

Of course, firms do not actually choose from a menu of budgets. However, they frequently can do what amounts to the same thing. They can impose budget overruns on government (taxpayers) or on consumers, if the environment is loose; that is, if competitive forces are inadequate, or

[2] While we carry on the analysis on the basis of individuals, we can with greater generality assume that some budget portions are allocated to small groups whose members' efforts are carried out jointly in some sense.

if regulations allow such overruns. At the very least budgetary permissiveness is likely to exist because of inert areas for consumers; that is, demand curves are likely to be bands rather than lines. Thus the size of the budget depends on environmental tightness (for example, options open to consumers depending on market structure, or degree of bureaucratic control over government subsidies, etc.).

The above leads to the basic proposition of X-efficiency theory: In a budgetary permissive environment the looser the ERC_j for all j on the average, the greater the degree of X-inefficiency (i.e., the excess of actual over minimum cost.)

V. Individual Firm Interactions and Environmental Tightness

The nature of the theory is sketched with the aid of three diagrams. In Figure 1, U_I is the relation between utility and effort for an isolated individual. The same utility/effort relation for the individual as a member of a team is U_T; it reflects the added satisfaction an individual gets by identifying with team effort. We can visualize entropy as an attenuation of team spirit, and a reduced identification with team results; hence a shift of U_T toward U_I.

In Figure 2 similar ideas from the viewpoint of individual-firm interactions are indicated. Pressure, which reflects the efforts of authorities as well as of peer groups, will help to determine individual performance. If pressure is less then

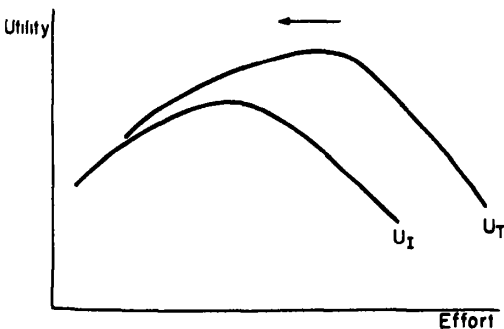



FIGURE 1

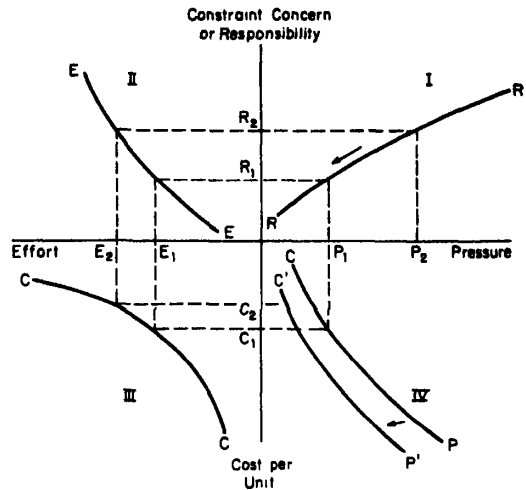



FIGURE 2

constraint concern is less, performance will be lower; and vice versa. The curve marked *RR* relates pressure "by the firm" on the individual, who responds by assuming a degree of constraint concern (i.e., responsibility). Pressure is the independent variable, constraint concern is the dependent variable. In quadrant II the curve *EE* relates constraint concern (now the independent variable) to exerted effort. In quadrant III the curve *CC* relates effort to that part of cost of production per unit contributed by the individual. A higher degree of effort is associated with a lower cost. Finally, in quadrant IV, cost is the independent variable, which in turn determines the pressure that the firm will put forth on the individual. Note the set of mutually consistent values; that is the values represented by the equilibrium point $P_1R_1E_1C_1$. The components of the "point" $P_2R_2E_2C_2$ are not consistent with each other, but if we start at P_2 and move counterclockwise through the four quadrants, the components gradually approach the equilibrium values $P_1R_1E_1C_1$.

Figure 2 suggests how the entropy forces operate. The arrow in quadrant I shows the direction of the entropy forces when pressure is greater than equilibrium pressure. Of course, this is related to the

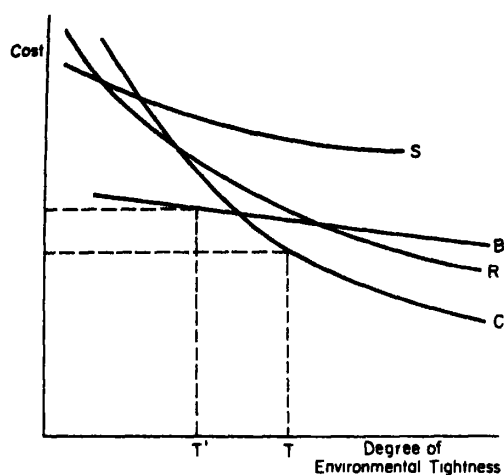



FIGURE 3

entire process of constraint concern effort-cost-pressure reaction, which leads to the movement along the curve *R*. However, there is another element involved; the movement of the cost-pressure reaction curve to the left—toward lower reaction levels. This is indicated in quadrant IV by the movement from *CP* to *C'P'*.

Four elements are likely to operate as cost containing factors, indicated by four curves in Figure 3. These are: standards of performance *S*, the maximum revenue obtainable *R*, bureaucratic controls *B*, and the degree of competition *C*. There are a variety of means through which the *tightness* of environmental cost-containing force manifests itself. For any degree of tightness *T*, the level of costs will be determined by the lowest cost-containment curve associated with the degree of environmental control. Thus in Figure 3 at point *T* the competition curve is the lowest cost-containment curve. Curve *C* is the operative control. As we shift to lesser degrees of environmental tightness other controls may become operative so that at *T'* the bureaucratic control is the binding constraint. In general, Figure 3 suggests the sources and levels of the supply of budgets in response to the demand for aggregate budgets discussed in Section III.

VI. Concluding Remarks

For a given output there is a budget supply determined by environmental tightness. There is a demand for budgets determined by the looseness of the *ERC* relations. In general, the looser the *ERC* relations, the greater the degree of X-inefficiency.

There are many areas (outside of perfect competition) where a theory of this sort might be applied. 1) In the case of regulated monopolies the fair profit rate on capital pricing rule does not protect the consumer. There is no motivation to minimize costs. 2) The production aspects of the nonprivate sector (public corporations, the post office, etc.) can be analyzed on the same basis. In this sector, adding funds need not add output proportionately, or add to output at all. It may only result in higher costs. 3) The supply of health services under a health insurance scheme is likely to create a means whereby the managers of "health firms" simply pass on higher costs to insurees. Hospitals may have an appetite for maximum budgets rather than to contain costs. 4) Similar analyses could be made of various municipal services, where the passing on of costs to the taxpayer may yield very high costs per unit of service. In other words, costs per unit would be determined by standards of performance or bureaucratic controls rather than through attempts at unit cost minimization.

I should note that the analysis does not necessarily imply tradeoffs between income and leisure when costs are higher rather than lower. The utility curve (Figure 1) which includes team satisfaction U_T can readily be higher than the one which excludes it, U_I. The entropy phenomenon which diminishes team participation satisfaction can reflect a movement toward lower utility levels.

REFERENCE

Harvey Leibenstein, *Beyond Economic Man,* Cambridge, Mass. 1976, chs. 5–11.

DISCUSSION

RICHARD H. DAY, University of Southern California: I come here in the role of discussant, not as critic but as advocate, not to bury but to praise. I applaud Harvey Leibenstein's attempt to formulate and apply realistic axioms of economic behavior; I appreciate Joseph Bower's focus on the actual structure of enterprise and am pleased to hear of this progress in showing how structure influences performance; and I declare an unabashed admiration for the logical clarity, good humoured ingenuity, and linguistic elegance of Sidney Winter's exposition of various central tenants of behavioral, evolutionary, or what I prefer to call *adaptive economics*. Some thirty years after Herbert Simon's early attempts to found it, a new field is clearly emerging. It may even have reached critical mass despite the reluctance of some of its most creative and in some cases vigorous exponents to admit that they belong to the same camp.

There are major problems with this new field, however: problems of formal definition; problems of empirical method; and problems of theoretical engineering. Just how bizarre some of these difficulties are can be grasped by contemplating some paradoxes, one of which is the fact that orthodox economists persist in using a patently false description of behavior and seem by doing so to possess a competitive advantage over those who use more realistic theory!

Economists on both sides, however, know the solution to the paradox, or should: optimality and equilibrium theory economize scarce planning and processing powers, just as myth, magic, and the analysis of dreams provided concise and plausible rationalization for our ancient predecessors when advising the governors of their day. Indeed, the orthodox economist's role is much like the shaman's of yore, often wrong, sometimes right, but usually when right, right for the wrong reasons.

The issue is confused by the necessity of us—as adaptive economists—to admit the inescapable relevance of the basic concepts of optimizing and equilibrium, even though in the admission we must demand a major reinterpretation. First, we must use traditional concepts to define what we mean by adaptedness or best economic fitness to the environment. Second, if we are to allow for learning in a state of incomplete knowledge, which is involved in a fundamental way in even the simplest gaming problems, then we must use optimizing with feedback, that is, local search of drastically simplified representations of the choice problem. In the form used by Richard Cyert and James March, learning is formally equivalent to a very simple recursive linear programming problem whose basic computation is the selection of the larger or smaller of two numbers after two elementary divisions. Of course, even this takes time, so learning (and economizing) in its crudest form cannot be the universal basis of behavior, though of course we hope it may be present in much of it.

To rephrase, learning involves optimizing not only in the sophisticated Bayesian sense, but also in the crudest behavioral, rule of thumb sense—and learning must be slow relative to the speed of many real processes because, even in its simplest form, it takes time. Since optimizing enters the scene as soon as we go beyond servomechanistic and other mechanical forms of behavior, we may see in it a kind of a priori concept, already existing in the brain—programmed there—to aid in the interpretation of and response to environment. But how different it is to think of optimizing as a broad class of adaptive rules, including a crude behavioral learning tactic, than to think of it as a definition and characterization of best states, acts, or rules.

Let me emphasize also the way facts are to be selected and used to develop adaptive theory. In dynamic economics we know that behavioral rules determine the qualitative character of trajectories. We can assume rational behavior in the orthodox sense and ask what parameters rationalize

histories. Great progress has been made by neoclassical theorists and econometricians in this vein. Or, we can model behavior using observed heuristics (involving mechanical rules, learning, or relatively simple optimizing) and see what are the consequences of each. Although this work began a century and a half ago (if we count Cournot), progress has been fitful and slow. The current set of papers is proof that some individuals are committed to advancing this cause even if in doing so they must admit their own ignorance. That admission, it seems to me, is a great step on the way toward a deeper understanding of the role played by intellect in the evolutionary process.

EFFECTIVENESS OF MONETARY, FISCAL, AND OTHER POLICY TECHNIQUES: COMPETING MEANS

What Can Stabilization Policy Achieve?

By ROBERT J. GORDON*

Skepticism about the role of discretionary or activist stabilization policy is not new, but within the past decade the balance of opinion in the economics profession has shifted sharply from widespread belief in the stabilizing potential of discretionary monetary and fiscal policy to a pandemic suspicion that such policy intervention may be incapable of yielding any net benefit. This paper briefly traces the intellectual and historical sources of this metamorphosis of opinion and attempts to reconstruct a qualified case for activism.

I. The Self-Assured Credulity of the Mid-1960's

At its zenith in early 1966, discretionary policy appeared to have achieved an unassailable victory over its critics. Few in the profession disagreed with Walter Heller's proclamation that "We now take for granted that the government must step in to provide the essential stability at high levels of employment and growth that the market mechanism, left alone, cannot deliver" (p. 9).

The theory of policy implicit in mid-1960's discussions called for maximization of an aggregate welfare function which depended 1) negatively on the absolute value of the "gap" between actual real *GNP* and the potential real *GNP* which could be produced at a 4.0 percent "full-employment" unemployment rate, 2) negatively on the inflation rate, and 3) positively on the growth rate of potential *GNP*. Maximization was to be performed subject to two major constraints: that the inflation rate depended negatively on the algebraic size of the gap; and that the nation's balance of payments could not be allowed to be too large a negative number.

Since the use of changes in government expenditures for stabilization purposes interfered with allocative considerations, frequent changes in income tax rates became the central policy tool. Monetary policy was kept in the background, relegated to the maintenance of a low and stable level of long-term interest rates to achieve the goal of stimulating potential output growth. The main loophole interfering with preference for tax changes was thought to be the legislative lag, and activist advocates tried without success to win approval for standby presidential authority to make quick temporary tax changes.

II. Elements in the Erosion of Support for Activism

A. *Forecasting and Lags*

Milton Friedman (1961) previously argued that long and variable lags in the effect of monetary policy were likely to make countercyclical monetary policy actions destabilizing. A subsequent theoretical analysis by Stanley Fischer and J. Phillip Cooper found that mere length of lags called for a *more active* policy keyed to the rates of change of target variables, but that variability of lags was a stumbling block which could well allow a Constant Growth Rate Rule (*CGRR*) for the money supply to outperform a more activist policy.

In the Fischer-Cooper analysis, policy

*Professor of economics, Northwestern University. This research has been supported by the National Science Foundation.

changes responded to variations in the actual values of a target variable and did not require any use of forecasts. Although the well-publicized failures of forecasters during the 1970's may appear to reinforce a skeptical disregard of forecast values, according to Stephen McNees, the record of four-quarter-ahead forecasts of real *GNP* during the 1970's actually was rather good, with the glaring exception of the special 1973–74 period dominated by unprecedented supply shocks. Further, both the length and variability of the lag in the effect of monetary policy may have been overstated. I have recently (1978) calculated that the average lag between the month of maximum monetary tightness and the subsequent onset of recession in four major post-Korean episodes was only 8.5 months, with a range between six and ten months.

B. *Uncertain Economic Structure and Policy Multipliers*

Present evidence provides no basis for confidence in the exact size of the impact of a policy change on target variables. In 1967 William Brainard showed that when policy multipliers are uncertain, the expected gap between actual and target *GNP* should be closed by only a fraction of the gap. Gary Fromm and Lawrence R. Klein, as well as Franco Modigliani and Albert Ando, exhibit widely varying estimates of both fiscal and monetary multipliers. Brainard's demonstration increases the danger that a policy stimulus introduced to close a *GNP* gap may lead to overshooting and an acceleration of inflation, or that policy restraint introduced to eliminate overheating will push the economy into a recession.

C. *The Natural Rate Hypothesis*

Milton Friedman's (1968) natural rate hypothesis (*NRH*) denied the ability of policymakers arbitrarily to select any inflation-unemployment combination along a stable tradeoff curve. Instead below a critical natural rate of unemployment the inflation rate would continuously accelerate, adding new urgency to Brainard's warning against overshooting the policy target. Some writers have denied the validity of the *NRH*, because they find unrealistic or unconvincing the classical equilibrium context in which its theoretical validity was demonstrated by Milton Friedman and others, with all economic agents on voluntary supply curves along which employment and output varied only if deviations between actual and expected price movements caused agents to be "fooled." However, Robert Barro and Herschel Grossman have shown that the *NRH* emerges also in a disequilibrium framework in which prices and wages respond to the excess demand for or supply of labor.

D. *The St. Louis Equation*

Soon after Milton Friedman's theoretical demonstration that the full-employment target of the activists might be unsustainable, Leonall Andersen and Jerry Jordan struck another blow with empirical equations which widened the range of previous multiplier estimates and implied that fiscal policy had no impact at all on nominal spending over as short a period as a year. Although activist advocates eventually regrouped and presented convincing evidence of fatal statistical flaws in the St. Louis procedure (see Alan Blinder and Robert Solow; Modigliani and Ando), their disarray lasted long enough partially to discredit fiscal activism and to allow the adoption by the Fed of monetary growth targets. Ironically, the gradual evolution of the data has steadily raised estimates of the St. Louis-type fiscal policy multipliers until recently they arrived in the vicinity of more conventional estimates (see Benjamin Friedman, 1977a). In the end, the St. Louis results served to stimulate useful analyses

of the theoretical conditions under which fiscal policy might have a zero multiplier in the long run, and also to shorten the consensus estimate of the lag of monetary policy.

E. *The Permanent Income Hypothesis and Temporary Income Tax Changes*

Robert Eisner, using Milton Friedman's permanent income hypothesis of consumption, showed that a temporary income tax cut or surcharge would fail to alter permanent income and thus would have a low spending multiplier. The temporary tax changes favored by mid-1960's activists were thus discredited as inappropriate for stabilization purposes, since their impact on consumption would not be large or rapid. Further, the lag in the effect of fiscal policy might be long and variable, with the length of the lag depending on the public's assessment of the likelihood that a tax change would soon be reversed.

F. *Reinforcement in the Late 1960's Policy Debacle*

These defects in the activist case might not have been so persuasive if they had not been accompanied by a remarkable coincidence of supporting events. Inflation accelerated between 1967 and 1969 far beyond the pre-1966 expectations of activist proponents. Further, inflation failed to slow down in the recession of 1970 and early 1971, as would have been expected along a fixed Phillips curve. The dramatic drop in the personal saving rate in late 1968 and the failure of spending growth to slow appreciably in response to the temporary tax surcharge was consistent both with the St. Louis claim that monetary multipliers had previously been underestimated and fiscal multipliers overestimated, as well as with the Eisner critique. Recent empirical work by William Springer and by Modigliani and Charles Steindel, reinforce the adverse verdict on the efficiency of temporary tax changes.

III. Rational Expectations, Supply Shocks, and Other Challenges of the 1970's

A. *Endogeneity of Structural Coefficients and Policy Multiplers*

Robert E. Lucas, Jr. added a new dimension to the Brainard analysis of policy multipliers by pointing out that both structural coefficients and policy multipliers were endogenous and would respond to the particular policies chosen, thus making the conduct of policy even more uncertain. For instance, workers have responded to the inflationary policies of the past decade by demanding much more complete indexation of wage contracts, thus altering the aggregate response of wage change to price change.

The insight that agents respond rationally to the policy environment has applied with special force to the recent behavior of financial markets. The short-run negative textbook correlation between the money supply and interest rates has been replaced by a positive correlation, as speculators observe the Fed's attempt to maintain monetary targets and bet that a high money supply outcome in a given week increases the probability that policy will be forced to shift toward restriction. But this new response pattern does not imply that an activist monetary policy is rendered impotent; market expectations are presently conditioned by knowledge that the Fed is attempting to pursue a particular target, and responses would change if the Fed were to alter that target to pursue a countercyclical activist policy stance.

B. *Application of Rational Expectations to Economic Policy*

Classical equilibrium versions of the *NRH* make changes in output depend on "surprises," that is, deviations between actual and expected prices. Thomas Sargent and Neil Wallace have argued that a monetary policy which reacts in a systematic way to past events, say a deriva-

tive control rule responding to past values of inflation and unemployment, cannot cause the required surprise, since rational agents will incorporate the systematic component of monetary behavior into their price expectations. Thus systematic countercyclical monetary changes can have no impact on real output, an apparently startling result.

Ironically, the Sargent-Wallace result, if true, would not only render policy impotent, but also make policy actions unnecessary. The price flexibility required to validate their result describes an economy with perfectly functioning self-correcting forces in which perceived shifts in aggregate demand alter prices but not output. In fact, in a Sargent-Wallace world the Fed could eliminate inflation simply by announcing that henceforth it would expand the money supply at a rate compatible with price stability. But today's world hardly appears consistent with the classical equilibrium interpretation of output fluctuations based on errors in forecasting prices. A large worldwide gap between actual and natural output has persisted in 1976–78 in the face of a relatively steady and well-predicted inflation rate.

C. *Supply Shocks and Legislated Inflation*

When a supply shock occurs, for example, the 1972–73 crop failures or the 1974 oil price increase, a *CGRR* policy condemns the economy to a simultaneous increase in both unemployment and inflation. The merits of an activist policy which increases the money supply to "pay for" the higher oil prices depends on the extent of wage indexation and the willingness of workers to accept a decline in the real wage (see the author, 1975). In retrospect, high unemployment in 1975–76 could have been substantially alleviated without an explosive inflation in the United States and Germany, but not in Italy and Britain, although the Lucas point requires this conclusion to be qualified for the possibility that American and German workers might not have been so docile under a more accommodative policy regime.

Prospective increases in payroll taxes, energy taxes, and the minimum wage in the late 1970's and early 1980's amount to a series of "mini supply shocks." A monetary authority adhering to a *CGRR* policy would find that these cost increases would raise unemployment. An accommodative monetary policy would shift the burden of the legislation from unemployment to real income losses for those holding assets yielding nominal-fixed returns.

IV. The Rehabilitation of Stabilization Policy

A. *Limitations of a CGRR Monetary Policy*

The existence of a potent self-correcting mechanism of price flexibility has again been refuted in the 1975–77 interval. Far from declining steadily and rapidly, the rate of change in *U.S.* wages has become stuck at a relatively constant rate since early 1976. Under these circumstances, a constant growth rate for the money supply cannot be appropriate. Ignoring changes in velocity, if the constant money growth rate is chosen as the current rate of inflation plus the natural (constant unemployment) growth rate of output, then the unemployment rate cannot fall. Only if inflation gradually abates can real output grow fast enough to allow unemployment to decline, but then a *CGRR* implies a steady shift in the composition of fixed nominal income growth from inflation to output growth, leading to a steadily accelerating expansion and inevitable overshooting of the target unemployment level. To avoid this, the constancy of money growth must eventually be abandoned. But if the sanctity of *CGRR* is to be violated in one direction, why cannot the pace of monetary growth be temporarily quickened in the early stages of the expansion to reduce the duration and extent of wasted resources?

An activist policy which concentrates its stimulus in periods when the economy is operating far away from target output es-

capes most of the problems raised by Brainard and Milton Friedman. Even the most radical proposals for monetary stimulus by activist advocates in 1975 called for elimination of only a fraction of the output gap in the first year. The trajectory of a two to three-year recovery implied in activist recommendations allowed plenty of time for adjustments to be made if multiplier estimates proved to be inaccurate.

The main disadvantage of activist antirecession monetary policy is that a temporary acceleration must inevitably be followed by a deceleration as the economy approaches its target. Political objections to requisite increases in interest rates, perhaps exacerbated by the proximity of an election, may hinder the "soft landing" approach to the target and lead to overshooting. But even then, the natural rate target is not a knife edge separating hyperinflation from hyperdeflation. Just as long-term labor contracts limit wage deceleration in recessions, so the acceleration of wages in an overheated economy is not instantaneous.

Even when the economy has arrived at its target output level, a *CGRR* is not appropriate. Benjamin Friedman (1977b) has extended William Poole's earlier analysis of a *CGRR* monetary policy, showing that the adoption of short-run targets for the money supply is efficient and correct only if the demand for money by the nonbank public is completely stable in relation to income and totally insensitive to interest rates. When the money supply grows less rapidly than expected in relation to income as in 1976, or more rapidly as in mid-1977, policy is efficient when it utilizes this information that the demand for money has shifted and deviates from its previous growth rate target.

B. *The Role of Fiscal Stabilization*

While Eisner's criticism of temporary income tax changes is convincing, insufficient attention has been given to other fiscal tools. In contrast to the income tax, temporary changes in subsidies and sales or payroll taxes are more effective than permanent ones by creating intertemporal displacement of spending. A reduction in a sales or payroll tax is exactly the opposite of a crop failure and allows policymakers to reduce unemployment and inflation simultaneously. Tax incentives for wage reductions also have this inflation-reducing beneficial impact. The main qualification is political; the necessity for congressional debate of such fiscal measures may lead not only to perverse spending effects in anticipation of future tax changes, but also to delay which causes tax changes to be made at the wrong stage of the business cycle.

The need to avoid political delay leads to renewed attention to automatic fiscal devices triggered by deviations of actual from target output. Exemplary applications abroad include the Swedish countercyclical investment fund, which allows corporations to escape tax on investment funds shifted from boom to recession periods, and the Japanese device of accelerating expenditures on public works in recessions (see the author, 1978, pp. 516–25).

The transition to flexible exchange rates in the 1970's has reinforced the case for using fiscal policy to stimulate the economy during recessions. While a monetary expansion boosts the supply of dollars and causes a *U.S.* exchange depreciation, fiscal policy raises the demand for money and appreciates the *U.S.* dollar. A stimulus which reduces the unemployment rate by a given amount will be accompanied by a stronger dollar, cheaper imports, and less inflation if it takes the form of fiscal rather than monetary ease.

V. Conclusion

Events in the late 1960's discredited the earlier brand of policy activism based on a permanent long-run inflation-unemployment tradeoff, and a hyperactive "fine tuning" technique.[1] But events in the 1970's

[1]Although the tax rebate of 1975 indicates that actions of politicians lagged behind the skepticism of economists.

support a reconstructed case for activism. When output is well below target, rigid adherence to a *CGRR* monetary policy leads to permanent acceptance of high unemployment if there is no downward adjustment of prices, and to overshooting the target if prices do adjust. Deviations from any reasonable estimate of target output have been large enough to allow a sizeable temporary stimulus without need for excessive concern about multiplier uncertainty. The experience of adverse supply shocks has focused attention on the potential role of subsidies, cuts in sales or payroll taxes, and wage-tax schemes as methods to achieve a simultaneous reduction in inflation and unemployment through the active use of policy.

While the economic case for stabilization policy seems convincing, political obstacles cannot be ignored. A temporary monetary stimulus in a deep recession may be well timed and effective, but the unwinding of the stimulus will require a politically unpopular increase in interest rates which the central bank may be forced to resist. Reductions in sales and payroll taxes may be easy for economists to recommend, but in actuality politicians are presently engaged in a major shift in the composition of federal tax revenue from the personal income tax to price-increasing payroll and energy taxes. One can only hope that Lucas' idea of policy-responsive parameters can be extended to the political sphere, and that politicians will learn from the sorry aftermath of their own current behavior to be less obstreperous in the future.

REFERENCES

L. C. Andersen and J. L. Jordan, "Monetary and Fiscal Action: A Test of Their Relative Importance in Economic Stabilization," *Fed. Reserve Bank St. Louis Rev.*, Apr. 1968, *50*, 11-24.

Robert J. Barro and Herschel Grossman, *Money, Employment and Inflation*, Cambridge 1976.

A. Blinder and R. Solow, "Analytical Foundations of Fiscal Policy," in *The Economics of Public Finance*, Washington 1974, 3-115.

W. Brainard, "Uncertainty and the Effectiveness of Policy," *Amer. Econ. Rev. Proc.*, May 1967, *57*, 411-25.

R. Eisner, "What Went Wrong?," *J. Polit. Econ.*, May/June 1971, *79*, 629-41.

S. Fischer and J. P. Cooper, "Stabilization Policy and Lags," *J. Polit. Econ.*, July/Aug. 1973, *81*, 847-77.

B. M. Friedman, (1977a) "Even the St. Louis Model Now Believes in Fiscal Policy," *J. Money, Credit, Banking*, May 1977, *9*, 365-67.

———, (1977b) "The Inefficiency of Short-run Targets for Monetary Policy," *Brookings Papers*, Washington 1977, *2*, 293-335.

M. Friedman, "The Lag in Effect of Monetary Policy," *J. Polit. Econ.*, Oct. 1961, *69*, 447-66.

———, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1-17.

R. J. Gordon, "Alternative Responses of Policy to External Supply Shocks," *Brookings Papers*, Washington 1975, *1*, 183-206.

———, *Macroeconomics*, Boston 1978.

Walter W. Heller, *New Dimensions of Political Economy*, New York 1966.

R. E. Lucas, Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Allan Meltzer, eds., *The Phillips Curve and Labor Markets*, Amsterdam 1976.

S. K. McNees, "The Forecasting Performance in the 1970s," *New England Econ. Rev.*, July/Aug. 1976; rev. 1977.

F. Modigliani and A. Ando, "Impacts of Fiscal Actions on Aggregate Income and the Monetarist Controversy: Theory and Evidence," in Jerome Stein, ed., *Monetarism*, Amsterdam 1976, 17-42.

——— **and C. Steindel,** "Is a Tax Rebate an Effective Tool for Stabilization Policy?," *Brookings Papers*, Washington 1977, *1*, 175-202.

W. Poole, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic

Macro Model," *Quart. J. Econ.*, May 1970, *84*, 197–216.

T. J. Sargent and N. Wallace, "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–57.

W. L. Springer, "Did the 1968 Surcharge Really Work?," *Amer. Econ. Rev.*, Sept. 1975, *65*, 644–59.

Labor Market Structure: Implications for Micro Policy

By CHARLES C. HOLT*

The current inability of economists to prescribe policies for achieving full employment without inflation can be traced to the unresolved schism between macro and micro analysis. There are calls for structural reforms that will impact at the micro level, but we still don't know enough about the microdynamics of the economy to specify effective programs. We need to identify the critical parameters at the micro level which account for frictional and structural unemployment and the bias toward inflation. This paper addresses one part of this problem by proposing an approach to explaining the dynamic and static structure of wages and unemployment in a segmented market, drawing policy implications and contrasting them with present programs.

I. Equilibrium Wages and Job Availability

Consider a labor market for a particular type of labor in partial equilibrium, but in addition to the usual allocation by real wages W, introduce allocation by "job availability" expressed in terms of the ratio of job vacancies to unemployment (V/U). A linear supply curve for labor is given by

$$(1) \qquad E_s = a + b\,W + c\,(V/U)$$

Employment supplied E_s is increased by higher wages or by greater availability of employment, that is, more unfilled job vacancies to be found in relation to the number of unemployed workers searching for them. Increased availability reduces average job search and waiting time with a corresponding increase of income and reduction of the psychic costs of unemployment thereby making employment more attractive relative to other activities.[1]

The demand by employers for workers E_d similarly is influenced both by wages and the unavailability of workers. High wage costs and additional high recruiting and training costs resulting from many vacancies relative to unemployed workers inhibit employment:

$$(2) \qquad E_d = d - e\,W - f\,(V/U)$$

where $(a, b \ldots f)$ are positive constants, and $(d > a)$ since the demand for labor exceeds the supply when $(W = V/U = 0)$.

Equating employment supplied and demanded and substracting (2) from (1) yield the equilibrium relation:

$$(3)\ (d - a) = (b + e)\,W + (c + f)\,(V/U)$$

which is shown in Figure 1. With two allocating variables, a set of equilibrium points can occur. Thus a unique equilibrium wage does not exist in this market.

Now consider segmented labor markets in which different types of labor are hired by different kinds of employers but compartmentalization is not complete, in that employers have some flexibility in hiring somewhat different but similar types of labor. As a result of local substitution, the equilibrium line (shown in Figure 1) for workers of type A overlaps in Figure 2 that for workers of type B and so on through the sequence of lower and lower paid types of labor. Local substitution between similar workers being paid similar wages will tend

*Director, Bureau of Business Research, University of Texas-Austin. I wish to acknowledge with thanks the helpful comments of Lorna Monti, Vince Geraci, Janet Barkley-Booher, and Camille Dvorsky.

[1]Many labor market decisions are known to be strongly influenced by the job availability variable. See references: Holt, Scanlon, Smith, Toikka and Vanski. It would be possible to lump "availability" into an adjustment on wages or alternatively into a job quality parameter but since the value of the (V/U) ratio is generated by market processes and operates as a second allocating variable in parallel with wages it is better to include it in the analysis explicitly.

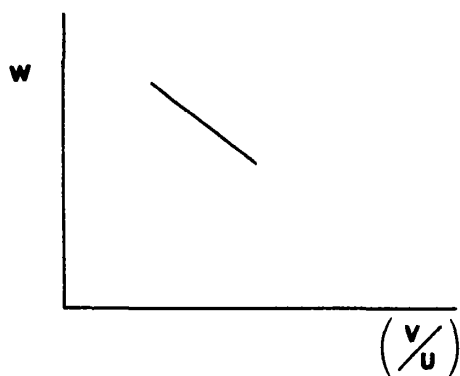



FIGURE 1. RELATIONSHIP OF WAGE AND JOB AVAILABILITY FOR A SINGLE SEGMENT OF THE LABOR MARKET

to yield the relation shown by the dotted line in Figure 2 that cuts across different types of labor. Relative wages ($W_i/\overline{W}$) are shown to depend on specific availability ratios $(V/U)_i$. A worker who could qualify for jobs in either segment A or B would tend to move to B because of the combination of better wages and/or better job availability. Employers that could recruit in segments A or B would tend to move to A because of lower wage cost and/or easier recruiting. The resulting supply shift from A to B and demand shift from B to A would raise the A equilibrium line and lower the B equilibrium line until they converged to the dotted line in Figure 2 representing the multisector

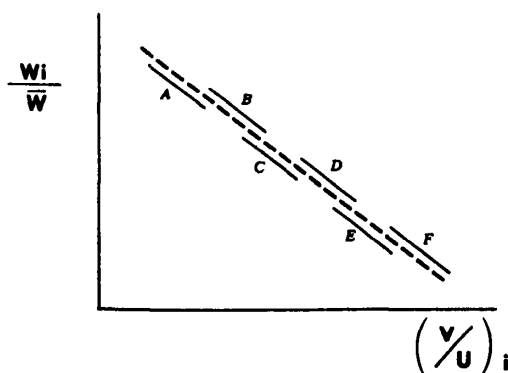



FIGURE 2. OVERLAPPING WAGE AND AVAILABILITY RELATIONSHIPS FOR ALL SEGMENTS OF THE LABOR MARKET

equilibrium. Of course, all other things are not equal so the relation will not be exact. Different types of workers and jobs will be found at different relative wages and job availability ratios, but there should be a systematic cross-section relation with negative correlation between wages and availability in equilibrium. To the extent that specific submarkets are compartmentalized without transfers between them, both wages and availability can be quite out of line in the interrelated markets of Figure 2.

The segmentation that separates different kinds of workers and jobs reflects basic skill and locational differences but in addition reflects differences in institutional access, information, and stereotyped employment roles associated with race, age, sex, culture, and tastes. Segmentation based on factors other than productivity affecting skills, of course, loosens the relation in Figure 2. The common sense of this important equilibrium relation is that high paying jobs are highly prized and many people direct their search toward them relative to the number of jobs available. In contrast, poor paying jobs are relatively abundant compared to the number of people who are looking for them.

II. Turnover and Equilibrium Wages, Employment, Unemployment, Participation, and Vacancies

I should stress that "availability" is a ratio. Although unemployment is high relative to vacancies at high wage rates, unemployment rates usually are high absolutely at low wage rates. This occurs because the quit plus layoff flows from employment are high at low wage rates. To offset the separations in equilibrium the hire flow must be correspondingly high and this requires relatively large stocks of vacancies and/or unemployed workers. Turnover rates will be relatively high where real wages and specific skills are low and where the ratio of vacancies to unemployment is high since quits are more responsive to market conditions than are layoffs. All of these influences contribute to high turnover

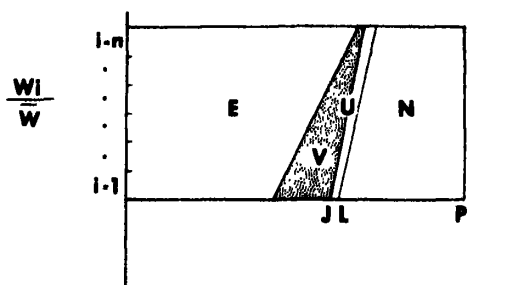



FIGURE 3. DISTRIBUTION OF EMPLOYMENT, VACANCIES, UNEMPLOYMENT, AND NONPARTICIPATION BY LABOR MARKET SEGMENTS

rates for low level jobs, causing correspondingly high levels of unemployment and/or vacancies.

Participation in the labor force is encouraged by high real wages, but is discouraged by high unemployment, since the drop out rate of the unemployed is much higher than for employed workers. Thus workers in low level jobs, other things equal, have relatively low labor participation because of low wages and high unemployment rates.

The wage-availability relation of Figure 2 interacts with turnover, drop out, and hire functions to determine in equilibrium the distributions of employment E, unemployment U, vacancies V (shaded), labor participation L, and out of the labor force N, across wage levels as shown in Figure 3. The working age population has been ordered by usual wage (or potential wage if individuals are not in the labor force) and put into n wage classes W_i of equal population size P. The average wage is $\overline{W}$.

While employers in response to the demand for goods and services can create jobs J, whether or not they are filled depends on decisions by people to forego leisure activities in N in order to participate in the labor force L, and to abandon the search for better jobs in U and accept employment in E. Thus equilibrium employment depends on joint decisions by workers and employers weighing wage considerations against the availability of alternatives in and out of the labor market.

The equilibrium is a stochastic one in which stable stocks are maintained by steady turnover flows and any deviations from the equilibrium state releases forces to restore it—provided the labor market segments are permeable. Because the annual flow through the labor market is nearly four-tenths of the force, the stock-flow equilibrium is established in a few months.

An examination of Figure 3 shows the results of interactions between wage and availability allocations, and turnover. Individuals who are perceived as having low productivity receive low wages. They experience unemployment frequently because they often quit and are laid off and even though they find jobs quickly their unemployment rates are high. That plus low wages make labor participation unattractive, so they often leave the labor force. The opposite generalizations apply to the individuals perceived as highly productive. Good jobs are highly sought after and relatively hard to find. However, once found, good jobs last a long time because of low quit and layoff rates, so unemployment is generally low for high wage workers.

III. Cyclical Changes in the Labor Market Equilibrium

When demand declines, parameter d in equation (2) falls, lowering the equilibrium relation in Figure 1 so that *both* wages and job availability are under pressure to decline.

As aggregate demand decreases, jobs are destroyed at all wage levels, but the workers who are forced into unemployment are most likely to be those at low wage levels. Employers attempt to keep workers with specific skills and they are likely to have higher wage levels. Many workers prefer to downgrade to lower wages or incomes rather than be laid off, thereby bumping lower paid workers. Union seniority also tends to focus the impact of layoffs on younger lower paid workers. The concentration of new unemployment at low wage levels puts maximum downward pressure on the lowest wage rates and increases the dispersion of relative wages.

Figure 4 illustrates the new equilibrium

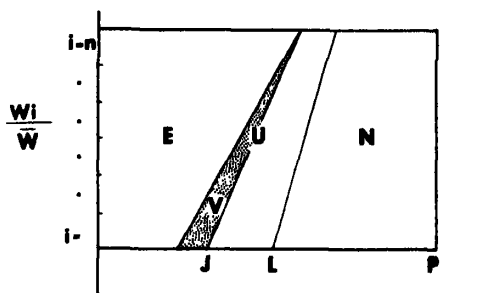



FIGURE 4. LABOR MARKET SEGMENTS UNDER DECREASED AGGREGATE DEMAND

with generally higher rates of unemployment but accentuated at the low wage end. Vacancies are generally low compared to the increased unemployment. Labor participation declines further so that only a fraction of the employment decline shows up as unemployment.

When aggregate demand increases, the opposite effects occur. Jobs are created at all wage levels and workers upgrade into better jobs leaving vacancies at the lower levels. The maximum wage impact is at the lower wage levels where the vacancies concentrate. As a result, wage dispersion declines in response to increased demand.

IV. The Wage Change Process

The foregoing analysis has been in terms of real wages, but for an inflation analysis we are concerned with changes in money wages. They tend to rise when 1) consumer prices increase leading to the decline of real earnings with benefit to profits, 2) the aggregate vacancy to unemployment ratio is high indicating generally tight labor markets, and 3) segments of the labor market are out of equilibrium—the excessively tight segments contribute more to wage inflation than the slack segments do to wage restraint.

When aggregate demand increases, employers raise prices and increase vacancies; the former lowers real wages and the latter increases recruitment costs. Workers respond to the former by reducing participation which tends to raise money wages while the increase in vacancies does that also. However, the wage adjustment process is very slow relative to the immediate increase in vacancies with a resulting large drop in unemployment. A reasonable Phillips curve adjustment would be:

$$\frac{\Delta w}{w} = g - h\left(\frac{w}{p}\right) + k\left(\frac{\overline{V}}{U}\right) + m\left(\frac{\Delta w}{w}\right)_{-1}$$

where w is the money wage rate, p is an index of consumer prices, and $(\overline{V/U})$ is the *average* vacancy to unemployment ratio. Disequilibrium in labor market segments would make this greater than simply the ratio of aggregate vacancies to aggregate unemployment. The last term introduces a distributed lag. The combined effect of raising vacancies and lowering real wages (hence the negative sign) constitute a measure of the aggregate demand pressure on the labor market and ultimately on the wage level. Vacancies can be eliminated by using the hire-turnover relation to obtain a more usual form for this relation, and a labor productivity trend should be introduced in the (w/p) term.

In spite of the fact that any increase in money wages tends quickly to be marked up in prices, which feed back into wages, the whole wage change process is very sluggish in response to changes in aggregate demand. Wages are almost fixed in the short run, so that during cyclical fluctuations, availability does most of the job of regulating the labor market decisions of employers and workers. However, hiring standards are rather quickly adjusted without changing money wages. This changes the price of labor expressed in productivity units, but there are practical limits to changing hiring standards before morale and productivity problems put a limit on this mode of adjustment.

Any competitive struggles for income shares between rival unions or between unions and employers with market power will put inflationary pressures on the levels of wages and prices. Also any increases in the relative prices of raw materials will cause temporary inflation until the adjustment is complete, and the inflation will be

increased if unions attempt to prevent the decline in real earnings.

V. Policy Implications

When aggregate demand is reduced in order to restrain inflation, the above analysis indicates that the greatest impact is in reducing the relative wages of the lower paid workers through substantial increases in their unemployment. This in turn triggers substantial labor withdrawal. As a result of the low skill level of these unemployed workers and their minimal substitutability for other workers, the resulting restraint on other wages is low. Thus the restriction of aggregate demand has a strong and highly inequitable impact on employment and earnings, and restrains inflation only weakly.

The injury to the most vulnerable workers by government use of aggregate demand restraint is reason enough for the government to pursue active measures at the micro level to minimize the causes of inflation and unemployment. The class conflict involved between those hurt by inflation and those hurt by the aggregate demand approach to restraining it should be avoided, if possible, by developing new policy alternatives.

Improving the functioning of the labor market at the firm and worker level can reduce unemployment and underemployment through better information, counseling and bringing workers and jobs closer to each other geographically, raising the quality of job-worker matches, increasing productivity through education, training and technology, and smoothing seasonal and other production fluctuations through better planning. The bias toward inflation can be reduced by decreasing the heterogeneity in educational and skill level, reducing power struggles for income shares, and reducing governmental regulations and licensing.

VI. Comparisons With Current Programs

Certain discrepancies between current and desirable policies are apparent.

Most manpower programs that were designed to improve the functioning of labor markets have been sacrificed for job creation. The result is that most manpower programs, as distinct from jobs programs, are funded at levels far below scales of efficient operation.

Efforts to promote the regional economic development of problem areas are in most states almost completely uncoordinated with manpower programs which are potentially highly complementary.

We have virtually no national and few state policies relating to the regional location of population and no mobility support programs which could help families solve their own employment problems through moving.

The knowledge base for increasing productivity and job satisfaction through matching peoples' preferences and abilities to jobs, training programs, and other services is seriously deficient because of the neglect by the government of behavioral research in this area. Even the academic knowledge that is available is little used in program operations.

Most manpower programs are targeted on the disadvantaged population or work primarily with low skilled occupations. As a result programs are perceived as offering little or nothing to employers or to fast moving skill bottleneck problems that contribute to inflation.

Compounding this problem is the close association, at least in the eyes of the employers, between the regulation and program functions of the Department of Labor.

The administration of programs still is seriously fragmented although progress is being made. The funding and administration of programs bearing on structural employment issues are split among six federal departments. The fragmentation of program qualification standards and administrative controls emanating from Washington still make it extremely difficult to achieve coherent program operations that address local needs. According to Niles Hansen, the A-95 process of local coordination of federal programs often

reaches program planning at stages that are too late to be effective. Lack of coordination with state programs raises additional problems. Federal technical assistance and training are inadequate. In spite all of these difficulties, a creditable job is being done.

Perhaps most critically there still is lacking the recognition by national policymakers that the problems of inflation and/or unemployment will persist until sufficient resources, leadership, and imagination effectively impact our neglected human resource problems at the firm and worker level.

The manpower dimension of our structural problem is, of course, not the whole story—capital, technology, resources, and international issues are also involved. But the economics profession must accept substantial responsibility for allowing the macro-micro split to become so large, and for counseling political leaders to restrain inflation through macro policies that cost hundreds of billions of dollars of lost output while neglecting to urge with equal cogency the pursuit of adequate and effective structural programs.

In the existing framework of economic analysis good policy prescriptions are nonexistent, as is clearly evident in current debates on economic policy. Given the economic structure, aggregate demand policies cannot achieve full employment without inflation. In the slack economy which results, the structural programs that are needed 1) encounter maximum political opposition because they affect distributional issues, 2) have reduced program impact, and 3) as a result, are funded far short of the level required to achieve significant structural reform.

These problems are now compounded by increased national concern for environment, energy, occupational health and safety, and employment discrimination which is expressed in new legislation, regulation, and litigation. These fall upon an economy in which demand has been made slack to fight inflation and adequate programs facilitating adjustments by employers and workers have been neglected.

Indeed there probably are no significantly better policies available than those currently being pursued in terms of *short-run* benefits since the economic structure cannot be improved quickly. The pitfall of present policies lies in their failure to recognize how much structure could be changed in the *long run* if structural policies were pursued consistently. The challenge facing economists is to formulate the analysis that would take into account the effect of aggregate demand on programs to improve economic structure at the micro level, and the potential fuller employment and reduced inflation which the improved structure would open for macro policy.

REFERENCES

Niles M. Hansen, *Improving Access to Economic Opportunity: Nonmetropolitan Labor Markets in an Urban Society*, Cambridge 1976.

C. C. Holt, "Modeling a Segmented Labor Market," in Phyllis Wallace, ed., *Women, Minorities, and Employment Discrimination*, 1977.

———, **R. E. Smith, and J. E. Vanski**, "Recession and Employment of Demographic Groups," *Brookings Papers*, Washington 1974, *3*, 737 60.

———, **R. S. Toikka, and W. J. Scanlon**, "Extensions of a Structural Model of the Demographic Labor Market," in Ronald G. Ehrenberg, ed., *Research in Labor Economics*, Vol. 1, 1976.

R. E. Smith, "A Simulation Model of the Demographic Composition of Employment, Unemployment, and Labor Force Participation: Status Report," in Ronald G. Ehrenberg, ed., *Research in Labor Economics*, Vol. 1, 1977.

R. S. Toikka, "A Markovian Model of Labor Market Decisions," *Amer. Econ. Rev.*, Dec. 1976, *66*, 821 34.

Efficient Disinflationary Policies

By ARTHUR M. OKUN*

The combination of exceptionally high inflation rates and unemployment rates has confronted *U.S.* policymakers with an unprecedented dilemma during the current expansion. They have responded with a compromise of sorts, aiming to achieve a gradual recovery in which unemployment rates inch back down to equilibrium over a prolonged period. I shall discuss the logic of the gradual-recovery strategy, and will outline an alternative, more efficient strategy of disinflation.

I. The Welfare Economics of Gradual Recovery

The gradual-recovery strategy has been enunciated by both the Ford and Carter Administrations. In January 1976, after the initial inventory snapback from the severe recession and with an unemployment rate of essentially 8 percent, the Ford economists drew a path to a 5.2 percent unemployment rate in 1980. Unemployment was thus to decline by 0.6 percentage point per year. This remains, more or less, the target path of the Carter Administration today. To me, it translates into a growth rate of real *GNP* of approximately 5.5 percent, taking 3.75 percent as the growth of potential *GNP* and assuming that a decline of 1 percentage point in unemployment is associated with 3 percentage points of extra real *GNP* relative to potential.

A hypothetical alternative strategy of strong recovery might have aimed at, perhaps, 7 percent growth of real *GNP*, and a decline in the unemployment rate of 1.1 percentage points a year, reaching the ultimate 5.2 percent in mid-1978. From 1976 to mid-1980, cumulative output along the gradual-recovery path is below that of the strong recovery path by about 10.5 percent of a year's *GNP*—a price tag of about $200 billion—ignoring the compound effects of the loss of physical and human capital. Similarly, the cumulated difference over the five-year period in the annual unemployment rate is 3.5 percentage points or "point years." The policymakers apparently judge that those costs of maintained slack are outweighed by its anti-inflationary benefits.

What is the disinflationary gain from the less rapid recovery? To evaluate this issue, I inspected six macroeconomic Phillips curves of recent vintage. (See Robert J. Gordon, Figure 3, p. 273; Robert Hall, Table 5, p. 378; Franco Modigliani and Lucas Papademos, Table 1, equation (1), p. 150; George Perry, Table 2, p. 416; James Pierce and Jared Enzler, equation 3, p. 19; Michael Wachter, Table 7, p. 146. Details of the calculations are available on request from the author.) While they are essentially accelerationist, implying no long-run tradeoff between inflation and unemployment, they all point to a very costly short-run tradeoff. For an extra percentage point of unemployment maintained for a year, the estimated reduction in the ultimate inflation rate at equilibrium unemployment ranges between one-sixth and one-half of 1 percentage point, with an average estimate of 0.3. Or, to put it another way, the average estimate of the cost of a 1 point reduction in the basic inflation rate is 10 percent of a year's *GNP*, with a range of 6 percent to 18 percent. The extra 3.5 point years of unemployment and the sacrificed $200 billion of output buys, according to these estimates, a reduction of between 0.6 and 1.8 points in the basic inflation rate for the 1980's.

II. The Costs of Anticipated Inflation

Is it worth paying $200 billion for 1 point or even 2 points of reduced inflation? I shall offer two contrasting answers to that question. On the standard view that anticipated

*The Brookings Institution. The views expressed are my own and are not necessarily those of the officers, trustees, or other staff members of The Brookings Institution.

(i.e., correctly predicted) inflation imposes no major social welfare costs, it is not worth anything approaching $200 billion.

In the standard formulation, an anticipated inflation rate of 6 percent is worse than a rate of 4 percent in only one respect: the extra resource cost of economizing on demand deposits and currency that bear zero interest in a world of higher nominal interest rates. And that welfare cost is trivial—surely not more than $1 billion a year.[1]

The view that the welfare costs of anticipated inflation are negligible is shared by some economists of various persuasions, but it is a central implication of the accelerationist-rational expectations school. In that model, a higher rate of anticipated inflation cannot raise output or employment, because economic behavior adjusts to it. Thus, it cannot do any good. But by the same token, since economic behavior does adjust to it, it cannot do any harm!

Of course, unanticipated inflation—any positive or negative deviation between the expected and realized inflation rate—has substantial welfare costs in these formulations. However it is a factual matter that inflation has been accurately predicted in 1976–77 by a broad consensus of professional forecasters, as well as by the bond market.

My own answer is different because I cannot accept the standard view. Contrary to its implications, people are disturbed by a high rate of anticipated inflation. After two years of well-predicted 6 percent rates of price increase, a majority of Americans name inflation as Public Enemy No. 1 repeatedly in opinion surveys. The American public cannot readily speculate, hedge, or arbitrage on inflation. Only a small percentage of workers have obtained effective escalator clauses on their incomes. Asset markets have not offered savers and investors a good hedge against inflation in the past dozen years—except for the illiquid and lumpy single-family home. Indeed, the most popular savings-type assets are yielding a negative before-tax real interest rate.

The specific ways in which people get hurt by inflation, even when they are not surprised by it, should be viewed in a broader context. Inflation disturbs an important set of institutions that economize on information, prediction, and transactions costs through ongoing buyer-seller relationships—what I have called customer product markets and career labor markets (1975). The bilateral monopoly surplus that develops in the interdependent economic relationships is preserved by accepted price or wage standards—implicit and explicit contracts, conventions, and habits—which are framed in currency units. These standards can adapt only slowly and painfully to inflation or slack. As a twin result, the short-run Phillips curve is remarkably flat, and even anticipated inflation is exceedingly costly.

I cannot hope here to win converts to this minority position on the welfare costs of inflation, which Sir John Hicks and I have taken, and which is consistent with the views Gardner Ackley presents persuasively elsewhere in this volume. But I would insist that every economist needs *some* rational explanation of why the current well-anticipated inflation is so disturbing to the public.

III. Slack as Insurance against Accelerating Inflation

Even those who deny the costs of anticipated inflation might espouse gradual recovery as a means of avoiding accelerated and unanticipated inflation if they saw much risk that a strong recovery would, in fact, raise the inflation rate well above 6 percent. But here again the standard formulation of a "natural" unemployment rate (or *NAIRU*) is reassuring, predicting that inflation will decelerate so long

[1]Suppose the elasticity of demand for M_1 with respect to nominal short-term interest rates is as high in absolute value as 0.3; then an interest rate of 8 percent, compared with 7 percent, would lower real money demand by 4 percent, or roughly $13 billion. The resource cost of more trips to the bank associated with that $13 billion differential might be about 7.5 percent of it, or nearly $1 billion.

as the unemployment rate remains above its equilibrium.

Unfortunately, history points to a less comforting verdict. Inflation has generally slowed or been reversed during recessions and the initial half-year (or perhaps year) of recovery. Beyond that, however, periods in which unemployment declined but remained above equilibrium have not typically witnessed decelerating inflation—not in 1933–37, nor in 1940–41, nor in 1961–64. (New Deal price-wage-cost policies simply cannot explain the paradox of the mid-1930's. Prices fell again in 1938–39 in response to recession, although the unemployment rate was lower than in 1934–35.)

The overall inflation rate, I submit, averages quick and slow responses to excess supply or demand. Auction product markets and casual labor markets respond very promptly. For example, wholesale prices of sensitive industrial materials *fell* 15 percent between May 1974 and March 1975, and then began rising. Such sectors add to inflation once the *GNP* gap starts narrowing, countering the lingering downward pressure on inflation from the customer and career sectors. Econometric Phillips curves that include as a variable the recent *change* in the *GNP* gap or in the unemployment rate illustrate this result. For example, Gordon finds that a 1 point *reduction* in the *GNP* gap adds as much to wage inflation as a gap *level* of 2.5 points subtracts from it (see his Table 3, column (8), p. 266).

The view that a full and strong recovery will not court additional inflation is optimistic in the light of history and especially optimistic in the present context of continued institutional adaptations to past inflation, rising relative prices of energy, and significant cost-raising measures taken by the government.

IV. The Cost-Reducing Strategy

In short, I believe it is important both to lower the current inflation rate and to ensure against a higher rate in 1978–79. Thus in my view—unlike the standard view—the strategy of maintaining slack is

not absurd. But it is inefficient. The efficient technique uses the direct influence of public policy on costs.

The basic analytics of the cost-reducing strategy can be seen in a simple accelerationist model in which the rate of increase of wages w in the current year depends on (a) the rate of price inflation p, of the last two years with coefficients summing to unity, and (b) the current rate of unemployment U. Those wage increases, in turn, feed into price increases fully with no lag. Thus,

$$(1)\quad w_t = \alpha p_{t-1} + (1-\alpha)p_{t-2} + f(U_t)$$

where $0 < \alpha < 1$

$$(2)\quad p_t = w_t$$

Now, suppose that equation (2) is disturbed by the introduction of a 1 percent subsidy on all items in the *GNP*, and that other fiscal and monetary actions are taken to hold the previously expected rate of unemployment. In the initial year, p will be pushed down by 1 percentage point (assuming, plausibly, full forward shifting of the subsidy). As a result, in the next year, w is lower by α. Since the second-order difference equation is dominated by a root of unity, both p and w are ultimately reduced by $1/(2-\alpha)$—a number somewhere between 0.5 and 1.0 (in percentage points). In this model, the 1 percentage point subsidy, enacted on a permanent basis, reduces the basic inflation rate about as much as two point years of extra unemployment does in the Phillips curve estimates cited above.

The hypothetical subsidy illustrates the character of cost reduction. Reductions in federal payroll taxes levied on employers and in state and local sales taxes are the closest analogies in the actual fiscal system. The cost-reducing strategy can also be pursued by subsidizing consumer goods with elastic supplies and inelastic demands, by designing farm-income supports that do not raise prices, by maintaining free access of imports, by bolstering the exchange rate of the dollar, and by relying on wage subsidies rather than minimum wages to aid low-income workers.

The simplified model exaggerates the disinflationary effectiveness of cuts in indirect taxes, relative to that of unemployment, through one of its features—the assumption that the feedback onto wages comes entirely from prices rather than from other wages as well. Any wage-wage feedback dilutes the effectiveness of cuts in indirect taxes, but does not alter that of higher unemployment. The wage-wage view is supported empirically by the 1973–74 experience, in which exploding fuel and food prices apparently did not add a great deal to *U.S.* wages, contrary to what a price-wage view would have predicted.

Suppose that, to allow for both wage-wage and price-wage feedbacks, equation (1) is replaced by

$$(3)\quad w_t = \beta w_{t-1} + \alpha' p_{t-1} + (1 - \alpha' - \beta) p_{t-2} + f(U_t)$$

Here, for the same permanent subsidy of 1 percentage point, the ultimate reduction in the inflation rate is $(1 - \beta)/(2 - \alpha' - \beta)$. In a process that is half wage-wage ($\beta = 0.5$), cuts in indirect taxes are, roughly speaking, one-third less effective than they are in a process that is purely price-wage. Obviously, if the feedback is entirely wage-wage ($\beta = 1$), then shocks that impinge directly on the price level have no lasting effect on inflation rates.

Thus, if the feedback process is mainly wage-wage, the cost-reducing strategy must get a direct handle on wage increases. Any wage-wage feedback must reflect a focus on *relative* wages. However relative wages cannot be altered by policy measures that raise disposable wage income generally, such as across-the-board wage subsidies or cuts in payroll taxes levied on employees. According to empirical investigations for the United States, cuts in the personal income tax do not slow wages, although that finding is questioned by some studies of other countries. Paradoxically, no general measure can break the wage-wage spiral in the same way that cuts in indirect taxes can break the price-wage spiral.

To break the wage-wage spiral, one must turn to penalties and incentives that alter the process of wage emulation. That is the basic analytical justification for various proposals that Henry Wallich and Sidney Weintraub, Abba Lerner, Laurence Seidman, and the author (1974, 1977) have advanced to slow down wage increases. Individual discretionary wage decisions have huge macro externalities. As the Phillips curve estimates suggest, an autonomous downward shift in the wage equation that produces a hold-down in wages of $1 permits an increase of roughly $6 in output, holding the inflation rate constant. In a sense, the social benefit of wage restraint in a slack economy is something like six times the size of the nominal gains forgone by the workers. No advocate of Pareto optimality should pass up such an opportunity for a deal!

So long as the link from wages to prices in the feedback (equation (2)) is reliable, a successful wage-slowing policy will reduce inflation. It is just as effective in a pure price-wage feedback system (like equation (1)) as in a wage-wage system. Finally, the implementation of any credible cost-reducing strategy should have additional favorable effects by lowering the inflationary expectations of well-informed observers, whose actions will then help to bring about the disinflation all the sooner.

All of the cost-reducing proposals are unconventional and unproven; many are inelegant and raise serious administrative problems. Those affecting wages ask business and labor to depart from their established patterns for maintaining career employment relationships. They introduce new elements into public finance choices that are already perplexing. But these difficulties should be weighed against the greatest inefficiency in our society—the waste of idle resources and the sacrifice of living standards and capital formation from maintained slack. The pursuit of efficiency calls for a major effort by the economics profession to design better disinflationary policies.

When the economy is plagued by inadequate or excessive demand, policymakers—sooner or later—apply the fiscal and monetary remedies that economists

have developed. The use of those remedies contributed mightily to the success story of the American economy in the 1950's and 1960's—a record of growth and stability unmatched in previous history, despite the inappropriate fiscal and monetary policies of two wartime periods.

In the 1970's, however, a new syndrome has emerged for which stimulus alone or restraint alone is not an efficient cure. A prolonged period of excess demand and upward cost shocks brought on the disease of inflationary momentum in our wage-price-wage feedback system. The cost-reducing strategy can cure that present problem efficiently, and it belongs on our shelf of countercyclical remedies for use whenever it is needed.

REFERENCES

R. J. Gordon, "Can the Inflation of the 1970s Be Explained?," *Brookings Papers*, Washington 1977, *1*, 253–77.

R. E. Hall, "The Process of Inflation in the Labor Market," *Brookings Papers*, Washington 1974, *2*, 343–93.

John Hicks, *The Crisis in Keynesian Economics*, New York 1974.

A. P. Lerner, "Stagflation—Its Cause and Cure," *Challenge*, Sept./Oct. 1977, *20*, 14–19.

F. Modigliani and L. Papademos, "Targets for Monetary Policy in the Coming Year," *Brookings Papers*, Washington 1975, *1*, 141–63.

A. M. Okun, "The Great Stagflation Swamp," *Challenge*, Nov./Dec. 1977, *20*, 6–13.

———, "Incomes Inflation and the Policy Alternatives," in *The Economists' Conference on Inflation: Report*, Vol. 1, Washington 1974, 365–75.

———, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers*, Washington 1975, *2*, 351–90.

G. L. Perry, "Determinants of Wage Inflation around the World," *Brookings Papers*, Washington 1975, *2*, 403–35.

J. L. Pierce and J. J. Enzler, "The Effects of External Inflationary Shocks," *Brookings Papers*, Washington 1974, *1*, 13–54.

L. S. Seidman, "A New Approach to the Control of Inflation," *Challenge*, July/Aug. 1976, *19*, 39–43.

M. L. Wachter, "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers*, Washington 1976, *1*, 115–59.

H. C. Wallich and S. Weintraub, "A Tax-Based Incomes Policy," *J. Econ. Issues*, June 1971, *5*, 1–19.

Unemployment Policy

By ROBERT E. LUCAS, JR.*

The *U.S.* unemployment rate was certainly too high in 1975, and most economists would agree that it is too high today. It will also be agreed that this observation poses a problem for public policy (in a sense that the observation that winters in Chicago are "too cold" does not). But what exactly is meant by the statement that unemployment is "too high," and what is the nature of the policy problem it poses? This question can be answered in more than one way, and the answer one chooses matters a great deal.

One common answer to this question is that there exists a rate of unemployment—call it "full employment"—which can and should serve as a "target" for economic policy. Unemployment above this rate is regarded as being of a different character from the "frictional" unemployment required to match workers and jobs efficiently, and is treated from a welfare point of view as waste, or deadweight loss. Elimination of this waste is an objective of monetary, fiscal, and perhaps other policies. In the first part of this paper, I will argue that this way of posing the issue does not lead to an operational basis for unemployment policy, mainly on the ground that economists have no coherent idea as to what full employment means or how it can be measured.

An alternative view, prevalent prior to the Great Depression and enjoying something of a revival today, treats *fluctuations* in unemployment and other variables as posing a policy problem. On this view, the average (or natural, or equilibrium) rate of unemployment is viewed as raising policy issues only insofar as it can be shown to be "distorted" in an undesirable way by taxes, external effects, and so on. Nine percent unemployment is then viewed as too high in the same sense that 2 percent is viewed as "too low": both are symptoms of costly and preventable instability in general economic activity. In the concluding part of this paper, I will sketch the approaches to unemployment policy which are suggested by this alternative view and some which are not.

I. Full Employment: Definition and Measurement

The idea that policy can and should be directed at the attainment of a particular, specifiable *level* of the measured rate of unemployment (as opposed to mitigating *fluctuations* in unemployment) owes it wide acceptance to John Maynard Keynes' *General Theory*. It is there derived from the prior hypothesis that measured unemployment can be decomposed into two distinct components: "voluntary" (or frictional) and "involuntary," with full employment then identified as the level prevailing when involuntary unemployment equals zero. It seems appropriate, then, to begin by reviewing Keynes' reasons for introducing this distinction in the first place.

Keynes (ch. 2, p. 7) classifies the factors affecting equilibrium employment in a real general equilibrium theory: the mechanics of matching workers to jobs, household labor-leisure preferences, technology, and the composition of product demand. Is it the case, he asks, that spontaneous shifts in any of these four real factors can account for employment fluctuations of the magnitude we observe? Evidently, the answer is negative. It follows that two kinds of theory must be needed to account for observed unemployment movements: granted that real general equilibrium theory may account for a relatively constant, positive component, *some other theory* is needed for the rest.

*University of Chicago. I am very grateful for criticism of an earlier draft by Jacob Frenkel, Sherwin Rosen, and Jose Scheinkman.

Accepting the necessity of a distinction between explanations for normal and cyclical unemployment does not, however, compel one to identify the first as voluntary and the second as involuntary, as Keynes goes on to do. This terminology suggests that the key to the distinction lies in some difference in the way two different types of unemployment are *perceived by workers*. Now in the first place, the distinction we are after concerns *sources* of unemployment, not differentiated types. One may, for example, seek very different theoretical explanations for the average price of a commodity and for its day-to-day fluctuations, without postulating two types of price for the same good. Similarly, one may classify motives for holding money without imagining that anyone can subdivide his own cash holdings into "transactions balances," "precautionary balances," and so forth. The recognition that one needs to distinguish among sources of unemployment does not in any way imply that one needs to distinguish among types.

Nor is there any evident reason why one would *want* to draw this distinction. Certainly the more one thinks about the decision problem facing individual workers and firms the less sense this distinction makes. The worker who loses a good job in prosperous times does not *volunteer* to be in this situation: he has suffered a capital loss.[1] Similarly, the firm which loses an experienced employee in depressed times suffers an undesired capital loss. Nevertheless the unemployed worker at any time can always find *some* job at once, and a firm can always fill a vacancy instantaneously. That neither typically does so *by choice* is not difficult to understand given the quality of the jobs and the employees which are easiest to find. Thus there is an involuntary element in *all* unemployment, in the sense that no one chooses bad luck over good; there is also a voluntary element in all unemployment, in the sense that however miserable one's current work options, one can always choose to accept them.[2]

Keynes, in chapter 2, deals with the situation facing an *individual* unemployed worker by evasion and wordplay only. Sentences like "more labor would, as a rule, be forthcoming at the existing money wage if it were demanded" are used again and again as though, from the point of view of a jobless worker, it is unambiguous what is meant by "*the* existing money wage." Unless we define an individual's wage rate as the price someone else is willing to pay him for his labor (in which case Keynes' assertion above is *defined* to be false), what *is* it? The wage at which he would *like* to work more hours? Then it is *true* by definition and equally empty. The fact is, I think, that Keynes wanted to get labor markets out of the way in chapter 2 so that he could get on to the demand theory which really interested him. This is surely understandable, but what is the excuse for letting his carelessly drawn distinction between voluntary and involuntary unemployment dominate aggregative thinking on labor markets for the forty years following?

It is, to be sure, possible to write down theoretical models in which households are faced with an "hours constraint" limiting the hours they can supply at "the" prevailing wage, and in which, therefore, there is a clear distinction between the hours one can supply and the hours one would like to supply. Such an exercise is frequently motivated as an attempt to "explain involuntary (or Keynesian) unemployment." This misses the point: involuntary unemployment is not a fact or a phenomenon which it is the task of theorists to explain. It is, on the contrary, a theoretical construct which Keynes introduced in the hope that it would be helpful in discovering a correct explanation for a genuine phenomenon: large-scale fluctuations in measured, total unemploy-

[1]Given the time-consuming nature of job search and the element of luck involved in finding a good "match," there is a capital-like element in most jobs. With job-specific human capital, the capital loss involved in job (or employee) loss is increased.

[2]These observations refer to easily verified features of any sizable labor market. Aggregate statistics on unemployment or on listed vacancies do not bear on their accuracy, since listing oneself as unemployed does not imply that one would accept *any* employment, nor is an advertised vacancy available to *any* job applicant.

ment. Is it the task of modern theoretical economics to "explain" the theoretical constructs of our predecessors, whether or not they have proved fruitful? I hope not, for a surer route to sterility could scarcely be imagined.

In summary, it does not appear possible, even in principle, to classify individual unemployed people as either voluntarily or involuntarily unemployed depending on the characteristics of the decision problems they face. One cannot, even conceptually, arrive at a usable definition of full employment as a state in which no involuntary unemployment exists.

In practice, I think this fact has been recognized for some time. Estimates of full employment actually in use have been obtained using aggregate information rather than data on individuals. As recently as the 1960's it was widely believed that there was some level of aggregate unemployment with the property that when unemployment exceeded this rate, expansionary monetary and fiscal measures would be noninflationary, while at rates below this critical level they would lead to inflation. One could then identify unemployment rates at or below this full-employment level as frictional or voluntary, and unemployment in excess of this level as involuntary. It was understood that only unemployment of the latter type posed a problem curable by monetary or fiscal policy. As Walter Heller wrote, "Gone is the countercyclical syndrome of the 1950's. Policy now centers on gap closing and growth, on realizing and enlarging the economy's non-inflationary potential" (Preface). Later, Heller refers to "the operational concepts of the 'production gap,' 'full-employment surplus,' the 'fiscal drag,' and 'fiscal dividends'" (p. 18).

For the purpose of calculating the production gap to which Heller referred, it makes little difference whether the voluntary-involuntary terminology accurately reflects differences in the way unemployed people view their situations. The issue here is rather whether there exists an aggregate rate of unemployment (on the order of 4 or 5 percent) which is of use in measuring an economy's noninflationary potential. If there were, then objections of the sort I have raised above could be dismissed as merely terminological: if one objected to calling unemployment above the designated full-employment level involuntary, one could call it something else, perhaps wasteful or unnecessary.

The last ten years have taught us a great deal about this operational concept of a production gap. In 1975, the *U.S.* economy attained the combination of 9 percent inflation and an unemployment rate of 9 percent. Applying the concept of a production gap to these numbers, does one conclude that the noninflationary potential of the *U.S.* economy is associated with unemployment rates in excess of 9 percent? Does one redefine 9 percent inflation to be noninflationary? Or can the entire episode be somehow pinned on oil prices?

I have reviewed two possible routes by which one might hope to give the term full employment some operational significance. One was to begin at the individual worker level, classifying unemployment into two types, voluntary and involuntary, count up the number classed as voluntary, and define the total to be the unemployment level associated with full employment. A second was to determine the operating characteristics of the economy at different rates of unemployment, and then to define full employment to be the rate at which inflation rates are acceptable. Neither of these approaches leads to an operational definition of full employment. Neither yields a coherent view as to why unemployment is a problem, or as to the costs and benefits involved in economic policies which affect unemployment rates. The difficulties are not the measurement error problems which necessarily arise in applied economics. They arise because the "thing" to be measured does not exist.

II. Beyond Full-Employment Policy

Abandoning the constraint that any discussion of unemployment must begin first by drawing the voluntary-involuntary distinction and then thinking in separate ways about these two types of unemployment

will, I think, benefit both positive and normative analysis. Practicing social science is hard enough without crippling oneself with dogmatic constraints. A terminology which precludes asking the question: "Why do people choose to take the actions we see them taking, instead of other actions they might take instead?" precludes any serious thinking about behavior at all.

Whether or not the body of work stemming from the Edmund Phelps volume, and earlier work of George Stigler, John McCall and others, has produced all the right answers about the determinants of employment and unemployment, it has at least begun to pose some of the right questions. By treating all unemployment as voluntary, this work has led to the examination of alternative arrangements which firms and employees might choose to adopt for dealing with fluctuations in product demand, and their reasons for choosing to react to such fluctuations in the way we observe them doing. Pursuit of this question has indicated both how very difficult it is, and even more so how much economics was swept under the rug by "explaining involuntary unemployment" by incompetent auctioneers or purely mechanical wage and price equations.

Practicing normative macroeconomics without the construct of full employment does take some getting used to. One finds oneself slipping into such sentences as: "There is no such thing as full employment, but I can tell you how it can be attained." But there are some immediate benefits. First, one dispenses with that entire meaningless vocabulary associated with full employment, phrases like potential output, full capacity, slack, and so on, which suggested that there was some *technical* reason why we couldn't all return to the 1890 workweek and produce half again the *GNP* we now produce. Second, one finds to one's relief that treating unemployment as a voluntary response to an unwelcome situation does not commit oneself to normative nonsense like blaming depressions on lazy workers.

The effect it does have on normative discussion is twofold. First, it focuses discussion of monetary and fiscal policy on *stabilization*, on the pursuit of price stability and on minimizing the disruptive effects of erratic policy changes. Some average unemployment rate would, of course, emerge from such a policy but as a by-product, not as a preselected target. Second, by thinking of this natural rate as an equilibrium emerging from voluntary exchange in the usual sense, one can subject it to the scrutiny of modern methods of public finance.

To take one example, as the level of unemployment compensation is varied, an entire range of average unemployment rates, all equally "natural," is available to society. At one extreme, severe penalties to declaring oneself unemployed could reduce unemployment rates to any desired level. Such a policy would result in serious real output losses, as workers retain poor jobs too long and accept poor jobs too readily. An output-maximizing unemployment compensation scheme would, with risk-averse workers, involve a subsidy to being unemployed, else workers retain a poor but relatively sure current wage in preference to the riskier but, on average, more productive return to seeking a new job. In view of the private market's inability to provide sufficient insurance against unemployment risk, still further gains in expected utility could be expected by still higher unemployment compensation, resulting in a deliberate sacrifice in real output in exchange for a preferred arrangement for allocating risk.[3] Notice that as one traces out tradeoffs of this sort, the issue of slack or waste does not arise. Different policies result in different levels of real output, but output increases are necessarily obtained at the expense of

[3]See Kenneth Arrow's analysis of medical insurance.

something else. Whether any particular level of unemployment compensation is too high or too low is a difficult issue in practice, but it is one that cannot be resolved simply by observing that other, unemployment reducing, compensation levels are *feasible*.

The policy problem of reducing business cycle risk is a very real and important one, and one which I believe monetary and fiscal policies directed at price stability would go a long way toward achieving. The problem of finding arrangements for allocating unemployment risks over individuals in a satisfactory way is also important, and can be analyzed by the methods of modern welfare economics. The pursuit of a full-employment target which no can measure or even define conceptually cannot be expected to contribute to the solution of either problem.

REFERENCES

K. J. Arrow, "Welfare Analysis of Changes in Health Coinsurance Rates," in Richard N. Rosett, ed., *The Role of Health Insurance in the Health Services Sector*, New York 1976.

Water W. Heller, *New Dimensions of Political Economy*, Cambridge, Mass. 1966.

John M. Keynes, *The General Theory of Employment, Interest, and Money*, London 1936.

J. McCall, "The Economics of Information and Optimal Stopping Rules," *J. Bus.*, July 1965, *38*, 300-17.

Edmund S. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1969.

G. J. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213-35.

The Invisible Fist: Have Capitalism and Democracy Reached a Parting of the Ways?

By SAMUEL BOWLES AND HERBERT GINTIS*

The twentieth century may be seen in retrospect as the era of the fruition and the collapse of liberal democracy. The extension of the suffrage and the advance of civil liberties is without a doubt one of the brilliant achievements of the late capitalist epoch, but it is a profoundly unstable achievement as well. The dynamics of liberal democratic capitalism have propelled us towards a fateful crossroads: one way—the extension of capitalism; the other—the extension of democracy. The choice itself heralds the twilight of the liberal tradition, which since the early nineteenth century has maintained the compatibility of capitalism and liberal democracy.

Late in his life, Jeremy Bentham joined with the philosophic radicals and drafted a Parliamentary motion advocating universal adult male suffrage. That this position marked a sharp break with early nineteenth century liberal opinion is suggested by the fact that Bentham felt compelled two years later, in 1820, to publish a tract reassuringly entitled *Radicalism Not Dangerous*, in which he disassociated the cause of suffrage from that of revolution, anarchism, and leveling. The movement for liberal democracy—that is, for an inclusive electorate and a roughly equal access to the public contestation of political issues, or what might alternatively be termed political equality and majority rule—dates from this period. So too does the thesis that liberal democracy and capitalism are uniquely compatible systems of allocation and decision making. This compatibility thesis has attracted distinguished advocates from Jeremy Bentham to Milton Friedman and has, over the past century and a half, enjoyed an impressive degree of popular endorsement.

The association of capitalism and liberal democracy in the popular mind may be traced to the historically coincident development of the two systems, and to the fact that no economic system other than capitalism has coexisted with liberal democracy. This is compelling testimony indeed. However the near universal assent to the compatibility thesis among liberal intellectuals is buttressed, understandably enough, by more theoretical arguments.

The celebrated parallelism of liberal democratic theory and *laissez-faire* economics would seem to assure compatibility. Both the economic and the political theory, the reader will recall, posit a society in which, with two significant exceptions, all important social relations are mediated through markets. Economic interdependence based on the division of labor is reconciled with individual autonomy through the working of the invisible hand. Thus all economic relations can be depicted as transactions for which contracts may at least in principle be written. Relations among family members of course constitute an exception, one which in the liberal economic and political tradition is neatly elided by rendering the concept "individual" indistinguishable in use from "family." The second exception is the relationship of citizen to the state. This relationship is mediated by competitive electoral mechanisms precisely parallel to those of the competitive market. The twin concepts of consumer sovereignty and citizen

*University of Massachusetts. An earlier version of this paper was read at the American Political Science Association annual meetings in Chicago in August 1976. Thanks to Ken Dolbeare, Chris DiStefano, Richard Edwards, and David Gordon for helpful criticism.

sovereignty, as well as the dual forms of political and economic competition, are based on legal equality, voluntary participation, and full information and appear to insure both the efficient and consumer-responsive allocation of economic resources and the popular accountability of the only major center of power, the state.

In early nineteenth century liberal thought, the economic basis for the presumed compatibility of capitalist economics and liberal democratic politics was the property-owning economy of yeoman farmers and other independent producers. Yet the introduction of wage labor, the *sine qua non* of capitalist production, does not by itself upset the theory. The competitive firm, as Paul Samuelson has taught us, is not a significant decision-making body; the capitalist is reduced by competition to the choice of obeying the market-enforced dictates of technical efficiency or going out of business.

This striking parallelism in liberal economic and political theory did not escape Karl Marx. Indeed, the compatibility of liberal political and economic institutions was strongly affirmed in Marx's work: the exploitation of labor and hence the generation of capitalist profits is perfectly consistent with formal legal equality, liberal democracy, and a system of markets which, he wrote, "represents a very Eden of the innate rights of man," a "realm of Freedom, Equality, Property, and Bentham." In the hands of Marx and Engels, Lenin, Luxemburg, and later theorists in this tradition, the compatibility thesis has, to be sure, constituted part of the critique of liberal political theory: the liberal democratic state, according to the *Communist Manifesto*, "is but a committee for managing the common affairs of the bourgeoisie as a whole." The Marxian critique has made clear the hiatus between political equality and majority rule as characteristic of a decision-making *process*, and popular sovereignty as an historical *outcome*. However, the rather compelling proposition that liberal democracy in the context of a capitalist economy does not insure popular sovereignty is a critique of liberal democracy itself, not of the thesis that capitalism and liberal democracy are compatible.

Diverse though it is, support for the compatibility thesis is hardly universal. From de Tocqueville to Dahl, liberal observers of American democracy have pointed to the possibility of contradictions between the capitalist growth process and the perpetuation of liberal democratic institutions. Some Marxists, in turn, have posited fascism as a likely outcome of the capitalist growth process. Prodded by recent events in Chile and by the rise of large left electoral parties in Europe, socialist scholars have begun to reconsider the relationship between liberal democracy and capitalism.

A second look at the historical record itself invites such a reassessment. Only a small minority of the capitalist societies may be termed even approximations of liberal democracy, less than a quarter by Dahl's count. Outside of Europe, North America, and the Commonwealth, there are less than half a dozen liberal democracies, down from about a dozen at the beginning of the previous decade. Further, the ills of the capitalist economy in the United States and Western Europe alike are increasingly attributed to what a report of the Trilateral Commission has termed "the excess of democracy."

These observations are clearly inadequate—for both logical and empirical reasons—to refute the compatibility thesis. However, they do cast doubt on any assertion of a necessarily positive relationship between capitalism and liberal democracy. An evaluation of the compatibility thesis itself requires more careful attention to the underlying theory.

We will argue that the compatibility thesis is false. Like the theory of perfect competition, it can be shown to be dynamically unstable. More technically we will suggest that the necessary conditions for the long-term reproduction of a liberal democratic capitalist society are contradictory. The internal dynamics of such a social order tend to undermine the necessary conditions for both capitalism and liberal democracy.

Three serious flaws in the compatibility thesis account for this dynamic instability. The first may be attributed to the instability of perfect competition itself. The competitive process has produced its own antithesis, the giant corporation. The associated concentration of control over the investment process, over the formation and articulation of public opinion, and more broadly over the development of policy alternatives and over political resources generally clearly violates the principle of roughly equal access to public contestation of political issues. Relations between the modern corporation and the rest of society are mediated neither by competitive markets nor by the liberal democratic electoral process. The corporation thus constitutes an unaccountable center of economic and political power in capitalist society.

The second flaw in the compatibility thesis concerns a contradiction between the necessary ideological conditions for the reproduction of liberal democracy and the conditions for the perpetuation of capitalist control over the production process. This flaw, like the first, may be traced to an error in the liberal economic theory.

The capitalist economy, even in its competitive form, cannot be adequately conceptualized as a system of voluntary exchange relations. The "realm of ... Bentham" does not encompass all or even the most central economic relations. What workers sell to the capitalist is not work itself—labor—but merely the formal jurisdiction over their capacity to work—labor power. However the profit of the capitalist depends not upon labor power, but labor itself—the concrete productive activity of work. While the buying and selling of labor power is a market-mediated relationship, the process of getting work out of the worker is not. Exceptional circumstances aside, the amount of work to be done cannot be expressed contractually, but rather is the outcome of an on-going and never completely resolved conflict between worker and employer. By erroneously considering labor itself as a commodity, liberal economics has overlooked the essential role played by the direct exercise of power by the capitalist over the worker. "Economics has earned the title of queen of the social sciences," aptly remarked Abba Lerner, "by choosing as its domain solved political problems." Contemporary liberal political scientists have shared the neoclassical economists' disinclination to explore the noncontractural relations central to the operation of the capitalist enterprise.

Earlier liberals, Adam Smith and Alexis de Tocqueville, for example, did entertain the notion that the division of labor within the capitalist enterprise might be antithetical to democracy. Yet the hierarchical control over the labor process, that is, over a substantial portion of the lives of most adults, does not in itself constitute a formal departure from liberal democracy. It does, however, give rise to a contradiction in the reproduction of a liberal democratic capitalist social order. From John Stuart Mill to Gabriel Almond and Sidney Verba, political theorists of diverse outlooks have affirmed that widespread democratic commitments, or more generally a democratic culture, is a necessary condition for the perpetuation of democratic government. In the advanced capitalist societies the discrepancy between a liberal democratic ideology and a daily life of hierarchical domination manifests itself in popular struggle to extend the sphere of democratic decision making to the labor process itself. However, widespread democratic contestation over the structure and control of the labor process would threaten the conditions for the reproduction of capitalism. Indeed the extension of anything more than token democracy to the workplace would most likely set in motion the progressive deterioration of not only capitalist control over production but of profits as well. Thus the democratic values essential to the reproduction of liberal democracy constitute an ideological climate which undermines the conditions essential to the reproduction of capitalism.

The third flaw in the compatibility thesis, like the second, is traceable to the liberal theorist's abstraction from the political conditions for the reproduction of the capi-

talist system. The question of how a liberal democratic political regime might be compatible with the reproduction of a class society has been explored, with varying intents, by writers as diverse as James Madison and Antonio Gramsci. We may posit two conditions under which such reproduction is facilitated: the existence of a multiplicity of politically relevant groups, classes, or strata, and the elimination from political contestation of issues which divide the political public among class lines. In most countries which are now liberal democracies, early competitive capitalism approximated these conditions. The coexistence of distinct forms of production—small farming, artisan production, capitalist production—and the imperfect regional integration of the economy gave rise to a large number of nonclass-related political groupings and minimized political polarization along the capital versus labor dimension. Further, the state, even where its economic stance was interventionist, was not directly involved in any substantial way in the wage labor—capital relationship. Class relations in production were for the most part mediated outside of the state sphere. Thus conflicts stemming from the direct exploitation of one class by another were defined as beyond the limits of political discourse and contestation. Indeed, the major economic intervention of the liberal state, tariffs, served to foster political unity among the producers of particular commodities (and consequently along regional or urban rural lines) rather than along class lines.

The very success of the capitalist accumulation process has radically undermined both conditions. The integration of national economies and the elimination of noncapitalist forms of production has reduced the importance of regional, peasant, and petty bourgeois political groupings. The capitalist class itself has been sharply reduced as a percentage of the population. Correspondingly the accumulation process has by any reasonable count created working class majorities in all of the advanced capitalist countries.

At the same time, the capitalist class in attempting to secure the conditions for the continuation of the accumulation process has conceded and even promoted a more direct intervention of the state in the wage labor—capital relationship. Further, the conflicts engendered by the accumulation process, and the market failures associated with urbanization, industrial concentration, the business cycle, and the like have prompted a partial supercession of markets as the primary allocational mechanism. The state has thus become a major economic actor directly involved in production, in the coordination of economic activity, in the mediation of class relations, and in the distribution of economic rewards. Class related issues have thus been rendered unambiguously political: class struggles have been partially displaced into the state sphere.

The implications for the long-term reproduction of liberal democratic capitalism are hardly auspicious. Popular movements around such issues as civil rights and ecology have imposed restrictions on the concept of private ownership of the means of production. The legal recognition of labor unions and the development of substantial systems of income support have dramatically altered the labor market and transformed the exchange of labor power for a wage into a quasi-political relationship. A significant portion of the customary standard of living of the working class—between a fifth and a third of the total wage bundle in all advanced capitalist countries except Japan—is now allocated through political mechanisms in the form of medical care, income support, public schooling and the like. Most significant, perhaps, democratically won gains in social expenditures—what *Fortune* charmingly terms "social drag"—appear to constitute a significant obstacle to the capitalist growth process. In the United States the net distributional impact of federal, state, and local taxes, transfers and expenditures appears to have been significantly and increasingly egalitarian over the years 1950–70. In the advanced capitalist countries as a whole, investment ratios and the ratio of social service expenditures to gross

domestic product have exhibited over the past two decades an increasingly strong negative relationship. Similarly, growth rates of total output and the ratio of social service expenditures to gross domestic product exhibit a strong negative relationship.

While these data are, of course, consistent with a variety of explanations, they are hardly supportive of the thesis of compatibility of liberal democracy and capitalism. Indeed they suggest that the growth process of capitalism, a necessary part of its reproduction, has generated a constellation of political forces which, while hardly revolutionary, appears to promote economic stagnation.

Much of the evolution of liberal social theory over the past century may be viewed as a response to these contradictions. Two strands are particularly striking. One is a natural extension of liberal democratic theory, known to economists primarily through the work of Joseph Schumpeter. Drawing on the earlier arguments of Mosca, Pareto, and others, Schumpeter invokes the putative decision-making incapacity of the electorate to reduce democracy to a competition among elites. Expertise replaces participation as the guiding principle in this new theory of democratic elitism. The task of the electorate is ratification, or the selection of a new group of experts.

The other strand is the enduring effort to arrive at a satisfactory definition of a private sphere of economic and social relations, a sphere exempt from democratic claims and beyond the legitimate intervention of the state. The accumulation process itself has not been entirely supportive of this effort, as the recent history of the family reveals. General intellectual developments have been equally unaccomodating. Unlike the Lockean theory of property rights, the utilitarian theory now dominant in the social sciences, ethics, and jurisprudence has proved a porous bulwark against the state. The theory of the second best, the social welfare function, and lump sum transfers are evidence that modern welfare economics has all but jettisoned *laissez-faire* and property rights. Indeed, responding to the social conflicts and irrationalities of advanced capitalism, economic theory has on balance rationalized rather than chastized the interventionist state.

The dynamic instability of liberal democratic capitalism is a contribution to the dissolution of the still dominant but moribund corpus of liberal social theory. It also represents a challenge: the construction of a more adequate conceptualization of the relationship between liberal democracy and capitalism. Two elements of such a new theory would appear to be essential.

First, the relationship between capitalism and liberal democracy is not a logical relationship in which liberal democracy may be inferred from the structure of the economic system. Rather, liberal democracy must be understood as a historically contingent outcome of a particular alignment of class and other forces generated in important degree by the capitalist accumulation process itself. With the possible exception of the United States, the achievement of liberal democracy, as distinct from republicanism, dates from the late nineteenth and the twentieth centuries, that is, from the period of the rise of the proletariat, not from the ascendancy of the bourgeoisie. Indeed, universal suffrage was never a project of the bourgeoisie, but rather a hard fought concession won by farmers and workers. It seems reasonable to conjecture, analogously, that a socialist democracy cannot be conjured up from the blueprints of workers councils and transformed property relations, but will arise, if at all, from the configuration of class forces which bring socialism into existence and impart to it its own dynamic.

Second, the extension of the suffrage and the securing of civil liberties during the capitalist era may best be understood as a political analogue to wage increases in the economic sphere. The exchange of labor power for the wage represents both a gain for the worker and a relinquishment. The receipt of the wage exhausts the worker's claim on the product and on participating in

the control over the production process. Historically the advent of the wage labor system represents the progressive elimination of a complex and varied system of claims on both the product and control of the process of production. In like manner, participation in the electoral sphere is a gain for the worker and a relinquishment. The worker gains a vote and the opportunity to contest in the electoral sphere. At the same time the worker gives up legitimate resort to other once very legitimate and effective forms of political expression: political strikes, bread riots, machine wrecking, the tarring and feathering of customs officers, and even dumping tea into Boston Harbor. The extension of the vote and the increased real wage are at once forms of integration of the working class and substantial working class gains.

If liberal democracy is not the logical political expression of the capitalist economy, if instead it is a bargain struck under duress, then the defenders of liberal democracy must strive continually to renew the terms of the accord. If our argument is correct they could hardly have chosen a less auspicious terrain upon which to take their stand than the advanced capitalist economy.

Indeed, there may be no viable defense strategy for liberal democracy. The successful mobilization of mass support for liberal democratic institutions may well entail a redefinition of democracy itself, one which emphasizes substantive popular outcomes as well as formal liberal democratic procedures. Thus the defense of liberal democracy may well entail more than the transformation of the capitalist economy. It may also set in motion forces for the supercession of liberal democracy itself, in favor of a socialist democracy which integrates political equality and majority rule with popular sovereignty.

Markets, States, and the Extent of Morals

By JAMES M. BUCHANAN*

Man acts within a set of institutional constraints that have developed historically: in part by sheer accident; in part by survival in a social evolutionary process; in part by technological necessity; in part by constructive design (correctly or incorrectly conceived). These constraints which define the setting within which human behavior must take place may, however, be inconsistent with man's capacities as a genuine "social animal." To the extent that moral-ethical capacities are "relatively absolute," (see Reinhold Niebuhr, pp. 3, 267), there may be only one feasible means of reducing the impact of the inconsistency. Attempts must be made to modify the *institutions* (legal, political, social, economic) with the objective of matching these more closely with the empirical realities of man's moral limitations.

In a certain restricted sense, the observed behavior of the modern American is excessively "self-interested." Rather than hope for a "new morality," I shall focus on the potential for institutional reform that may indirectly modify man's behavior toward his fellows. Institutions may have been allowed to develop and to persevere that exacerbate rather than mitigate man's ever present temptation to act as if he is an island, with others treated as part of his natural environment. In a properly qualified sense, the latter pattern of behavior is the economist's "ideal," but the costs have not been adequately recognized.

Let me proceed by simple illustration. Consider two traders, each of whom is initially endowed with a commodity bundle. Gains from trade exist and cooperation through trade is suggested, but there arises the complementary conflict over the sharing of net surplus. As we extend the model by introducing additional traders, however, the conflict element of the interaction is squeezed out, and, in the limit, each trader becomes a pure price-taker. "In perfect competition there is no competition," as Frank Knight was fond of emphasizing. (However, we must never lose sight of the elementary fact that this "economic ideal," including its most complex variants, presumes the existence of laws and institutions that secure private property and enforce contracts.)

Let me change the illustration and now assume that the same two persons find themselves in a genuine "publicness" interaction. (They are villagers alongside the swamp, to use David Hume's familiar example.) As before, there exist potential gains from trade, and these can be secured by agreement. Cooperation and conflict again enter to influence choice behavior, but here the introduction of more traders does nothing to squeeze down the range of conflict. Indeed, it does quite the opposite. Beyond some critical limit, each person will come to treat the behavior of others as part of the state of nature that he confronts as something wholly independent of his own actions.

Numbers work in oppositing directions in the two cases. Under a set of laws and institutions that are restricted to the security of property and contract, the extension of the market in partitionable goods moves the efficiency frontier of the community outwards. But, under the same laws and institutions, if there exist nonpartitionable interdependencies (public goods), an increase in the size of the group may move the attainable efficiency frontier inwards.

I have introduced the familiar private goods-public goods comparison to illustrate my general argument to the effect that there are opposing behavioral implications involved in any extension in the membership of a community. The effects of group size on choice behavior, and, through this, on the normative evaluation of institutions.

*Virginia Polytechnic Institute. I am indebted to Roger Congleton, Thomas Ireland, Janet Landa, Robert Tollison, and Richard Wagner for helpful suggestions.

have not been sufficiently explored by economists, most of whom have remained content to concentrate on the formal efficiency properties of allocations. With relatively few exceptions they have worked with fixed-sized groups. And even in 1978, most economic policy discussion proceeds on the implicit presumption that "government" is benevolently despotic.[1]

What is the orthodox economists' response when pure public goods are postulated? It is relatively easy to define the formal conditions that are necessary for allocative efficiency, but it is not possible to define the governmental process that might generate these results.[2] Work in public choice theory has contributed to our understanding of how governmental processes actually operate, but this theory is, in a general sense, one of governmental failure rather than success.

Political scientists have objected to the imperialism of public-choice economists who extend utility-maximizing models of behavior to persons who act variously in collective-choice roles, as voters, as politicians, and as bureaucrats. These critics intuitively sense that a polity driven solely by utility maximizers (with empirical content in the maximand) cannot possibly generate an escape from the large number analogue to the prisoners' dilemma suggested in the simple example of a public goods interaction. These critics have not, however, understood the basic causes for the general dilemma that modern collectivist institutions impose on citizens, politicians, and bureaucrats. Even more than the economists, orthodox political scientists have tended to ignore the possible effects of group or community size on individual behavior patterns.

Any political act is, by definition, "public" in the classic Samuelsonian sense. An act of voting by a citizen potentially affects a result that, once determined, will be applied to *all* members of the community. Similarly, an act by a legislator in voting for one tax rule rather than another becomes an input in determining a result that will define the environment for all members of the polity. Comparable conclusions extend to each and every act of a civil servant and to each decision of a judge (see Tullock). Under what conditions could we predict that such political acts will provide public good? For instruction here, we can return directly to our elementary example. We should expect at least some such behavior to exhibit cooperative features in effectively small groups. We should not, and could not, expect persons who act politically to provide public good voluntarily in large number settings.

We can reach this conclusion by economic analysis that incorporates standard utility-maximizing behavior on the part of all actors. My principal hypothesis, however, involves the possible inconsistency between man's *moral* capacities and the institutions within which he acts. Is not man capable of surmounting the generalized public goods dilemma of modern politics by moral-ethical principles that will serve to constrain his proclivities toward aggrandizement of his narrowly defined self-interest? It is here that my secondary hypothesis applies. The force of moral-ethical principle in influencing behavior is directly dependent on the size of community within which action takes place. Other things equal, the smaller the number of persons with whom a person interacts, the higher the likelihood that he will seem to behave in accordance with something akin to the Kantian generalization principle: in our terminology, that he will provide public good in his choice behavior.

Even this secondary hypothesis can be discussed in a way as to bring it within a utility-maximizing framework. The extent that a person expects his own behavior to influence the behavior of those with whom he interacts will depend on the size of the

[1] Economists have continued for eight decades to ignore the warnings of Knut Wicksell.

[2] A possible qualification to this statement is required with reference to the demand-revealing process, summarized by T. Nicolaus Tideman and Gordon Tullock. Even its proponents recognize, however, that this process remains a conceptual ideal rather than an institution capable of practical implementation.

group. Hence, utility maximization in a small number setting will not exhibit the observable properties of utility maximization in a large number setting (see the author). I want, however, to go beyond this strictly small group phenomenon of direct behavioral feedback. I want to introduce moral ethical contraints in a genuine noneconomic context here. I propose to allow "Homo economicus" to exist only as one among many men that describe human action, and, in many settings to assume a tertiary motivation role.

The precise dimension of human behavior that I concentrate on here is the location of the effective mix between the two motivational forces of economic self-interest and what I shall term "community."[3] I do not want to , and I have no need to, identify with any particular variant of nonself-interest: fellowship, brotherhood, Christian love, empathy, Kantian imperative, sympathy, public interest, or anything else. I want only to recognize the existence of a general motive force that inhibits the play of narrowly defined self-interest when an individual recognizes himself to be a member of a group of others more or less like himself. Robinson Crusoe could be motivated by nothing other than self-interest until Friday arrives. Once he acknowledges the existence of Friday, a tension develops and Crusoe finds that his behavior is modified. This tension exists in all human action and observed behavior reflects the outcome of some resolution of the inner conflict. The institutional setting determines the size of community relevant for individual behavior. This influence of size is exerted both directly in the sense of limits to recognition, and indirectly in the relationship between a community's membership and its ability to command personal loyalties. Conceptually, the "structure of community" within which an individual finds himself can shift the location of behavior along a spectrum bounded on one extreme by pure self-interest and on the other by pure community interest within which the actor counts for no more than any other member.

The institutions (economic, geological, legal, political, social, technological), which define the sizes of community within which an individual finds himself, impose *external* bounds on possible behavior. Parallel to these external constraints there are also *internal* limits or bounds on what we may call an individual's moral-ethical community. There are, of course, no sharp categorical lines to be drawn between those other persons whom someone considers to be "members of the tribe" and those whom he treats as "outsiders." I make no such claim. I assert only that, for any given situation, there is a difference in an individual's behavior toward members and nonmembers, and that the membership lists are drawn up in his own psyche. This is not to say either that persons are uniform with respect to their criteria for tribal membership or that these criteria are invariant with respect to exogenous events. Clearly, neither of these inferences will hold. However the fact of behavioral discrimination is empirical and subject to test. I am not arguing normatively to the effect that individuals should or should not discriminate among other members of the human species, or even as between humans and other animals.

My colleague Tullock enjoys asking egalitarians whether or not they would extend their precepts for social justice to the people of Bangladesh. He gets few satisfactory answers. Why should precepts for distributive justice mysteriously stop at the precise boundaries of the nation-state? If one responds that they need not do so, that national boundaries are arbitrary products of history, then one is led to ask whether or not effective precepts of justice might stop short of such inclusive community, whether or not the moral-ethical limit for most persons is reached short of

[3]In a tautological sense, all behavior, including that which I label as moral-ethical, can be analyzed in a utility-maximizing model. In this paper, however, "utility maximization" and "self-interest" are defined operationally.

the size of modern nations.[4] At provincial or regional boundaries? At the local community level? The extended family? The clan? The racial group? The ethnic heritage? The church membership? The functional group?

What can a person be predicted to do when the external institutions force upon him a role in a community that extends beyond his moral-ethical limits? The tension shifts toward the self-interest pole of behavior; moral-ethical principles are necessarily sublimated. The shift is exaggerated when a person realizes that others in the extended community of arbitrary and basically amoral size will find themselves in positions comparable to his own. How can a person act politically in other than his own narrowly defined self-interest in an arbitrarily sized nation of more than 200 millions? Should we be at all surprised when we observe the increasing usage of the arms and agencies of the national government for the securing of private personal gain?

The generalized public goods dilemma of politics can be kept within tolerance limits only if there is some proximate correspondence between the external institutional and the internal moral constraints on behavior.[5] This century may be described by developments that drive these two sets of constraints apart. An increase in population alone reduces the constraining influence of moral rules. Moreover population increase has been accompanied by increasing mobility over space, by the replacement of local by national markets, by the urbanization of society, by the shift of power from state-local to national government, and by the increased politicization of society generally. Add to this the observed erosion of the family, the church, and the law—all of which were stabilizing influences that tended to reinforce moral precepts—and we readily understand why "Homo economicus" has assumed such a dominant role in modern behavior patterns.[6]

Indirect evidence for the general shift from morally based resolution of conflict and morally based settlement of terms of cooperation to political-legal instruments is provided by the observed rapidly increasing resort to litigation. Modern man seeks not to live with his neighbor; he seeks instead to become an island, even when his natural setting dictates moral community. This movement, in its turn, prompts lawyers to turn to economic theory for new normative instruction.

Despite the flags and the tall ships of 1976, there is relatively little moral-ethical cement in the United States which might bring the internal moral-ethical limits more closely in accord with the external community defined inclusively by the national government. There is no "moral equivalent to war," and, since Viet Nam, we must question whether war itself can serve such a function. Nonetheless, experience suggests that war and the threat thereof may be the only moral force that might sustain the governmental leviathan. Viewed in this light, it is ominous that each president, soon after entering office, shifts his attention away from the divisive issues of domestic politics toward those of foreign affairs. We must beware the shades of Orwell's *1984*, when external enemies are created, real or imaginary, for the purpose of sustaining domestic moral support for the national government.

While I am not some agrarian utopian calling for a return to the scattered villages on the plains, I shall accept the label of a constitutional utopian who can still see vi-

[4]In an argument related to that in this paper, Dennis Mueller concentrates on the relationship between the size of community and the ability of a person to imagine himself behind a Rawlsian veil of ignorance.

[5]Gerald Sirkin refers to the "Victorian compromise" which is, in several respects, similar to the correspondence noted here.

[6]My diagnosis is restricted to the Western, specifically the American, setting. Perhaps the strongest empirical support for my argument is, however, provided in non-Western collectivized countries through the observed failures to create "new men" via institutional change.

sions of an American social order that would not discredit our Founding Fathers. To achieve such an order, drastic constitutional change is surely required. Effective federalism remains possible, within the technological constraints of the age, and "constitutional revolution" need not require the massive suffering, pestilence, and death associated with revolution on the left or right. Dramatic devolution might succeed in channelling some of the moral-ethical fervor in politics toward constructive rather than destructive purpose.

I become discouraged when I observe so little discussion, even among scholars, of the federal alternative to the enveloping leviathan. Where is the Québec of the United States? Where is the Scotland? Could a threat of secession now succeed? More importantly, could the emergence of such a threat itself force some devolution of central government power? Who will join me in offering to make a small contribution to the Texas Nationalist Party? Or to the Nantucket Separatists? From small beginnings . . . .

We should be clear about the alternative. The scenario to be played out in the absence of dramatic constitutional reform involves increasing resort to the power of the national government by those persons and groups who seek private profit and who are responding predictably to the profit opportunities that they observe to be widening. Individually, they cannot be expected to understand that the transfer game is negative sum, and, even with such understanding, they cannot be expected to refrain from investment in rent seeking. Furthermore, as persons and groups initially outside the game come to observe their own losses from political exploitation, they too will enter the lists. As the process moves forward through time, we can predict a continued erosion of trust in politics and politicians. But distrust will not turn things around. "Government failure" against standard efficiency norms may be demonstrated, analytically and empirically, but I see no basis for the faith that such demonstrations will magically produce institutional reform. I come back to constitutional revolution as the only attractive alternative to the scenario that we seem bent to act out. In the decade ahead, we shall approach the bicentenary of the Constitution itself. Can this occasion spark the dialogue that must precede action?

REFERENCES

J. M. Buchanan, "Ethical Rules, Expected Values, and Large Numbers," *Ethics*, Oct. 1965, *74*, 1 13.

D. Mueller, "Achieving the Just Polity," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 147-52.

Reinhold Niebuhr, *Moral Man and Immoral Society*, New York 1932.

G. Sirkin, "Resource X and the Theory of Retrodevelopment," in Robert D. Leiter and Stanley J. Friedlander, eds., *The Economics of Resources*, New York 1976, 193–208.

T. N. Tideman and G. Tullock, "A New and Superior Process for Making Social Choice," *J. Polit. Econ.*, Dec. 1976, *84*, 1145–160.

G. Tullock, "Public Decisions as Public Goods," *J. Polit. Econ.*, July/Aug. 1971, *79*, 913–18.

Knut Wicksell, *Finanztheoretische Untersuchungen*, Jena 1896.

Illusions of Necessity in the Economic Order

By ROBERTO MANGABEIRA UNGER*

Three experiences of the unavailability of reform seem to draw an iron circle around the American economy and to establish a gruesome brotherhood of disappointment between would-be iconoclasts and despondent technocrats.

The first of these impressions of intractability has to do with beliefs about the relationship between ideals of social life and the institutional contexts in which those ideals can be effectively realized. It is a widely held conviction that the world of work cannot or should not reflect the aspirations of democracy and community whose force Americans recognize in other areas of their lives. Any attempt to project these ideals into offices, shops, and factories will be both self-defeating and inefficient. There is a secret code of the moral life that decrees democratic conflict for mass politics, communal solidarity for the family, and unsentimental moneymaking or stoical discipline for exchange and work. To confuse the categories is to violate the code. And the sanction is this: that the very ideals whose range of application you set out to extend will evaporate or rot in your hands. You'll waste away your moral capital by investing it in aspects of society with a low rate of moral return, or else, like Midas, you'll be nonplussed by your own success.

The second face of necessity is that the desperate facts of luxury and indigence appear more likely to be changed, if they can be changed at all, by national emergency, general enrichment, or haphazard transfers than by any concerted design of redistribution. A vast net of mutual patronage, dependence, and threat links, in all directions, every sector of the state to every group in society. Any redistributive program that is more than a merciful tinkering with this chain begins to seem like a revolutionary fantasy. It will take something less whimsical to make the clear eyed and the comfortable lose any sleep over universal suffrage.

The third and most remarkable encounter with the unavoidable goes to the limits of control and foresight in economic policy itself. The puzzles of inflation, unemployment, and fiscal crisis are taken either as punishment for an enormous failure of self-restraint such as will repeatedly occur in the course of democratic politics or else as embedded in market structures and habits of behavior that no one seems powerful enough to change. Those riddles seem traceable to the natural vices of a country where free men try to have their cake and to eat it too, where everyone wants to keep a step ahead so as not to fall a step behind, and where self-denial is a mark of stupidity, servility, or saintliness.

Each of these three images of necessity is at once the acknowledgment and the distortion of a truth. The three may seem far removed from each other. Yet, they are bound closely together and, in fact, bound together in a way that carries a message about the failures of economics as a science. My strategy will be to approach this trinity of fates—the impossibility of realizing democratic and communal ideals in the system of production, the impossibility of achieving fundamental redistribution, and the impossibility of managing economic disorder—from its third and most practical side. Only then will my argument take me to the other members of the trinity and to the nature of economics itself. Thank goodness that I have plenty of time and that we all have open minds.

The deep factors that limit government's ability to avoid the disorganization of exchange show up most prosaically and most stubbornly in the problem of inflation. To begin to comprehend these factors is more than to move beyond monetarist and

*Harvard University.

fiscalist, Phillips curve and natural rate of unemployment theories. It is also to reject as partial and therefore misleading those theories of noncompetitive behavior and nonclearing markets that try to build a half-way house between conventional economic analysis and the interpretation of a broader social and moral reality. For what existing economics is entitled to say about inflation consistently with its own methods and assumptions is like the thirteenth chime of a clock, which not only startles us but casts doubt on the previous twelve chimes and, indeed, on the clock itself.

The social basis of inflation can be pictured as a set of interrelated tensions. They all grow out of a historical experience in which hierarchies of position and advantage, inherited and acquired, remain real while also losing some of their coercive force and patina of sanctity. These stresses form the elements of a distinct style of corporatist politics and economy.

The first tension lies in the interplay between the invidious comparisons people make about each other's situations and the struggle in which they engage, according to the measure of their power, to preserve certain customs of stability and fairness. Everyone compares his own benefits with those of the better off, and this remorseless contrast becomes the great factory of needs and of diligence. Yet the all out contest for advantage that would put every social relation up for grabs, for the sake of short-run self-interest and allocational efficiency, is repeatedly deflected by another force. This is the effort to maintain stable ongoing relationships of employment and exchange, to distinguish them sharply from casual deals in labor or goods, and to inform them with customary standards of relative wages and prices, power and prestige, discretion and deference.

The second tension describes a strange blend of weakness and strength in the hold that government has over society. For one thing, the state lacks the effective power or the moral authority to champion new society-wide differentials of advantage or even to win allegiance to the established hierarchies and distributions. For another thing, however, every class and institution depends for position and advancement upon some form of governmental support—so that the pillaging of the state, the time-honored mark of aristocratic polities, now becomes a universal practice.

The third tension goes to the relationship between individual interests and the organizations that claim to speak on their behalf. The explosion of interests and needs in the delayed wake of the weakening and demoralization of class and communal order has a paradoxical effect upon political struggle. These somewhat inchoate and largely inarticulate aspirations tend to cluster around the parties, professional associations, and labor unions that can define them in the least divisive and most straightforward material terms. On the other hand, however, the relative plasticity of people's ambitions and solidarities enable the representative organizations to get a fix on these solidarities and ambitions: to freeze them into a mold that offers stability to the organizations themselves and that allows the leadership of each of them to deal with constituents, clients, partners, and governments in a way that leaves the rules of the game unchanged. The organizations and the half-mute concerns they supposedly represent end up hostage to each other, and the stranglehold produced by their reciprocal adaptation drastically reduces the freedom of maneuver available to any alternative kind of leadership.

Together these tensions produce a characteristic situation of blockage on the transformative uses of power and conspire to recreate in every sector of the society a politics of frantic bargaining and overall drift. The consequences of this predicament for inflation are these.

The government, facing groups whose political influence is only randomly related to their productive contributions, will be repeatedly tempted to finance its programs through means other than outright taxation or transfer and to emphasize those aspects of distribution and redistribution that are not directly connected with a plan of institutional change rather than those that are.

Workers will resist the disruption of their

customary wage structure; they will undermine national incomes policies either because these policies jeopardize the integrity of the wage structure in a particular segment of the workforce or because they shore up a society-wide pattern of distribution that is itself viewed as illegitimate; and their unions will turn away from a politics of broader conflict about institutions that seems to endanger the established habits, prerogatives, and personnel of the unions themselves. In the absence of any prospect of major shift of power or wealth, the fight to keep ahead of the others will absorb all attention, and all larger solidarities will dissolve into smaller ones.

Managers and capitalists will exploit every opportunity to expand their margins of profit and therefore of safety against the uncertainties of corporatist politics. They will defend the price and wage rigidities that, together with the exploitation of a perennially jobless underclass and with the exaction of favors from government, help underscore their stable links with their own core markets and workforce.

The corollaries of the blockage revealed in relatively harmless fashion in inflation reach into every aspect of society. They make the madness of war look like the only solvent of routine politics. They make it impossible to determine what people would in fact say and choose in a situation in which the blockade on reform were partially lifted. They lead the ideological imagination of the elites to exhaust itself in hesitancy between a half-hearted state paternalism and a pseudo-restorationist program of free competition, both incapable of changing the facts against which they rail.

Faced with this petrification of possibilities, the subversive political mind will track down, under the monolith of paralyzed power, the hidden fault lines of disbelief and contention. It will look for issues that are at the periphery of accepted political discourse and take their ambiguities to reformist advantage: the widening of the range of debate about the fiscal and regulatory links between governmental power and private privilege, which can be either an occasion to correct minor abuses or the beginning of a larger quarrel about the tie between state and society; the promotion of meritocratic opportunities in career advancement, with its uncertain bearing upon the need to create the starting points of equality from which the meritocratic race is to be run; the idea of workers' participation, which can be either a refined Taylorism or a step toward workers' control.

The subversive political mind will identify and seek implausible ties among groups that are both continuously created by the existing order and continuously frustrated by it: the manual underclass, locked into a cruel ghetto of the labor market; the white-collar workers, finding that they are divided from their blue-collar counterparts more by style and expectation than by power, income, opportunity, or even work conditions; the restless, fancy, ingenious staffer class, suffering from a chronic shortage of political place.

The subversive political mind will find ways to reconcile tactless visionary leadership with the tactful contrivance of new organizational bases of power and group alliances.

It is in the course of alliance making and ideological conversion that the relation of corporatist politics to the problems of redistribution and ultimate ideals becomes clear. The conditions under which the stalemate of corporatist politics can be broken are the only circumstances in which basic redistributive programs can be accepted and the institutions of work remade in the light of principles of democracy and community. At the same time, these principles and programs are the only impulse that can give, in the final analysis, coherence and momentum to the reconstructive effort. But, though it may be easy to understand the meaning of redistribution, however great the difficulties of bringing it about, it is hard to picture with precision and realism just what it might mean to bring community and democracy to the workplace and, indeed, to do so in a way that permanently extirpates the extremes of poverty and wealth.

Sometimes it is said that any talk of community and democracy in the organization

of work will be punished by a decline in productive efficiency. The plain fact of the matter is that no one can tell in advance how technical and political aims might be accommodated when neither the institutional forms of work nor the human responses to them can be taken as fixed points. Moreover, the present system of production, like every system of production, is a power order as well as a technical solution to the problem of scarcity. In the interest of its own preservation, it does and it must put brakes on that constant, flexible interplay between abstract task definitions and concrete operational experiences that is the very essence of our modern conception of efficient rationality, and it does so by distinguishing more or less sharply the people responsible for defining productive tasks from the people charged with carrying them out.

It is also claimed that there would be a fatal instability in any attempt to realize democracy and community in the workplace, that worker's democracy would produce a new panoply of hardened vested interests, and that worker's community could only spell a preposterous village tyranny of the workplace. But, for workers' democracy to take root, the institutionalization of conflict and the subordination of technical advice to collective choice at the workplace level has to be paralleled by an enlargement of democracy in the nation. Without expanded conflict and participation in national government and in planning or regulatory agencies, workers' control becomes the occasion for another politics of corporatism when it does not remain a cover up for bureaucratic manipulation. Conversely, without democracy in the enterprise, the despotism of private power and technical expertise over everyday life remains unbroken. It is essential to a democratic production system that the relationship between the enterprise and each level of government be regulated not by the allocation of absolute property rights, but by the distribution of complex entitlements of dissent, participation, and access to jobs, markets, and capital. This means that, as capitalist property rights in large-scale enterprise are phased out, rights of decentralized decision making and market exchange are acquired by workers themselves. But it also means that limits are set on the distribution of productive capital as wages, on the size of enterprises, and on the extent to which profits and investments can be used to control other people's labor in other firms.

Again, community in the setting of work must signify the democratization of mutual burdens and vulnerabilities as a basis for shared concerns if it is not to come down to an attempt by bosses or bureaucrats to ennoble power through a pretense of togetherness.

A society that has ceased to deal with the problem of power through the allocation of absolute property rights and that has rid itself of its obsession with the contrast between contract and community is not a society that merely projects existing communal and democratic ideals into unfamiliar terrain. For when the ideals come unstuck from their traditional contexts, they must be re-imagined. The realms of democratic conflict, communal solidarity, and technical hierarchy will overlap rather than oppose each other once they have been allowed to find new meanings and applications.

Economics as it now exists is incapable of understanding either the deep realities of the corporatist style of politics and economy or the conditions and consequences of its reformation. Its inadequacies in this regard come from two sources: the nature of its method as a science and the character of its practical relation to power.

The crucial point about method is what to make of the struggle for generality and formality in economics in the course of its passage from classical political economy to the marginalist theory of general equilibrium and from there to a still more overarching theory of maximizing behavior. It is a mistake to attack the science that has emerged from this line of development because it relies on simplifying assumptions or reaches counterintuitive results or searches relentlessly for new modes of formal analysis. All these may be among the wounds that bold and rigorous thought is entitled to inflict upon the preju-

dices of common sense for the sake of surprising insight. If it is true that this formalized economics was forged partly in response to the socialist attack on political economy, it is also true that it partly succeeded in creating a science that can be used by anyone whose ambitions and curiosity are sufficiently modest.

The central vice in this neoclassical theory lies rather in having been seduced and corrupted by too primitive a conception of rational order. This fall into methodological sin manifests itself in two complementary ways.

First, it shows up in the inability to carry through a sustained interplay between theoretical analysis and empirical discovery. As long as the actual complicated facts of work and exchange are an occasion for relaxing independently held assumptions or for judging departures from preconceived standards of allocational efficiency rather than a subject for cumulative causal theorizing, economics as a science is condemned to eternal infancy. It will be made up of explanatory schemes that are relatively immune from the interesting paradoxes of the historical world and of factual observations that are random and *ad hoc* with respect to the theoretical machines into which they are fed. Such a science cannot grow, and its precocious virtuosity in formalization must be followed by a terrible sterility of substantive insight.

The reverse side of the same failing lies in a central ambiguity that runs through the related concepts of maximization, efficiency, and rationality. Is there a hollow and tautologous notion of maximizing rational decision that means no more than pursuing your aims whatever they are in the world whatever it's like? Or must we read into the concept of rationality a much narrower set of assumptions about the free-flow of resources, the determinants of conduct, and the translatability of inequalities of power and disagreements of purpose into the language of material exchange? If the former, economics cannot rise above tautology except by the dangerous and undisciplined expedient of relaxing assumptions. If the latter, however, economics is tied down to constricting normative and empirical premises from whose hold it had rightly tried to escape and with whose truth it is congenitally unprepared to deal.

The task, then, is to imitate classical political economy in this respect: that theories of exchange must be folded back into theories of power and perception.

The other major way in which economics has become an obstacle to understanding and to change can be inferred, once again, from the contrast with classical political economy. The heart of the program of the classical economists lay neither in a mere hope of collective material improvement nor just in an attachment to individual autonomy, but rather in a vision of historical opportunity. The relative withdrawal of class and corporatist privileges appeared to leave an open space in society, a space in which reason and labor could reach new heights of invention. Even when the classical economists toadied to the great, their ideas had power and authority because these doctrines were inspired by a generous image of unrealized human possibility. In its acceptance of this inspiration, political economy was not confusing description and evaluation; it was tacitly acknowledging that to understand a form of life is to prefigure its hidden possibilities of transformation.

The open space on which the classical economists pinned their hopes seems much smaller to us than it seemed to them; and the area left vacant by the crack in class coercion and institutional discipline to be densely overlaid by a new fatalism of stalemate, distrust, and despair. Without a true practice of science and a transformative imagination, economics must kneel down before this latter-day apparition of fate as a reality that it cannot change and a mystery that it cannot grasp. Despite the seeming agnosticism of its methods and assumptions, economics will then become a metaphysic for solid operators, a coat of armor for the prudent and the passive, and a hocus-pocus in the vestibule of power.

CHANGES IN CONSUMER PREFERENCES

Endogenous Tastes in Demand and Welfare Analysis

By ROBERT A. POLLAK*

Economists have traditionally been suspicious of changing tastes, and a profession's intellectual tastes change slowly. Nevertheless, it is time to reconsider the conventional wisdom that tastes are none of our business.

Those who favor incorporating taste formation and change into economic analysis fall into two groups whose intersection is almost empty. One is primarily interested in the welfare implications of changing tastes, the other in the analysis of household behavior. John Kenneth Galbraith (1958, chs. 10, 11) provides an articulate statement of the welfare view. "Radical" economists of both Marxist and non-Marxist persuasions also reject the notion that the formation of tastes is outside the province of economics or else deny that economics can be separated from the other social sciences (see, for example, Herbert Gintis). But the recent impetus to incorporate taste formation and change into economic analysis has come primarily from those interested in household behavior rather than welfare, and the principal focus of this work has been empirical demand analysis. In Section I, I discuss the substantial progress which has been made recently in incorporating changing tastes into both theoretical and empirical demand analysis.

Taste formation and change pose more difficult problems for welfare analysis. Variable tastes undermine the normative significance of the fundamental theorem of welfare economics which asserts (in a precise sense and under fairly stringent assumptions) that in competitive equilibrium everyone gets what he wants, subject to the constraints imposed by technology, resources, and the satisfaction of the wants of others. However if tastes are sufficiently malleable, then this may be no more than a corrollary of the more general proposition that people come to want what they get. I discuss the welfare issues briefly in Section II.

*University of Pennsylvania. This research was supported in part by grants from the National Science Foundation.

I. Taste Differences and Taste Change in Demand Analysis

Taste differences among households associated with differences in their demographic characteristics have long been a mainstay of the analysis of household budget data, but specifications of taste formation and taste change have only recently come to play a role in the analysis of time-series data. The economist's traditional reluctance to investigate taste change is well expressed by Milton Friedman. After discussing the relative nature of human wants, he writes:

> Despite these qualifications, economic theory proceeds largely to take wants as fixed. This is primarily a case of division of labor. The economist has little to say about the formation of wants; this is the province of the psychologist. The economist's task is to trace the consequences of any given set of wants. The legitimacy of and justification for this abstraction must rest ultimately, in this case as with any other abstraction, on the light that is shed and the power to predict that is yielded by the abstraction. [p. 13]

Although Friedman expresses the dominant view, three points should be noted. First, his strictures apply to taste formation and taste change, not to taste differences: thus, the use of demographic characteristics in the analysis of household budget data does not violate his dictum. Second, Friedman recognizes that tastes are not really fixed and, by implication, that they are endogenous to the socioeconomic system. Nevertheless, he is willing to forego descriptive accuracy, arguing that taste formation and change are not the economist's business and that their study should be left to the psychologists. However, he regards the proper test of the validity of this division of labor as its power to predict. This is an essential point, because the recent impetus to treat taste formation and change has come largely from empirical demand analysis. Finally, Friedman's argument presupposes that we are exclusively concerned with "positive economics"; whether this division of labor between economics and psychology is appropriate for welfare analysis is a distinct issue which Friedman's argument does not address.

George Stigler and Gary Becker appear to take the more extreme position that economic analysis should not only shun taste formation and change but also taste differences. They rightly object to invoking taste differences and taste change as a *deus ex machina* to "explain" whatever we cannot otherwise explain; but they do not take an equally critical view of attributing observed differences in or changes in behavior to unobserved differences in or changes in household technology. Whether one accounts for the observation that "...exposure to good music increases the subsequent demand for good music..." (p. 78) in terms of taste change (as I would) or in terms of the accumulation of music capital (as Stigler and Becker do) is a matter of semantics, not substance. There is no more explanatory power in a household production model which postulates the accumulation of a specific but unobservable "consumption capital" (for example, "music capital," p. 79) than in a habit formation model of the type proposed by Richard Stone (1954, 1964a,b), Hendrick Houthakker and Lester Taylor, the author (1970), or Louis Phlips (1972).

There are a number of ways to incorporate taste change into the analysis of household behavior. The initial choice is between specifications which make taste change exogenous to the socioeconomic system (for example, a time trend on some of the parameters of the demand equations) and those which make it endogenous. I shall discuss two specifications of endogenous tastes, habit formation, and interdependent preferences.[1]

Habit formation can be incorporated into demand analysis by postulating that a household's preferences depend on its own past consumption.[2] I shall discuss only the "one-period lag" specification of habit formation (i.e., tastes depend only on con-

[1]Two other types of endogenous tastes are those influenced by advertising or by prices. Galbraith has emphasized the ability of producers to manipulate consumers through advertising. Characterizing his own work, he writes, "The surrender of the sovereignty of the individual to the producer or producing organization is the theme, explicit or implicit, of two books, *The Affluent Society* ... and *The New Industrial State*" (1970, p. 471). For a discussion of advertising and references to the literature, see Richard Schmalensee (ch. 4). Preferences for goods may depend directly on prices because people judge quality by price or because a higher price enhances "snob appeal." For a discussion of the implications of price dependent preferences and references to the literature, see the author (1977).

[2]Stone (1954, p. 522) first suggested that habit formation could be incorporated into demand analysis by allowing some of the parameters of the linear expenditure system to depend on past consumption; he implemented his own proposal in his 1964a,b articles. Houthakker and Taylor proposed and estimated a model in which past consumption influenced present consumption patterns through a "state variable" which they interpret as a "psychological stock" of habits. In their second edition, they showed how their dynamic demand functions could be obtained from utility maximization. Various theoretical models of habit formation are investigated in W. M. Gorman; Maurice Peston; the author (1976b); Carl von Weizsäcker; Wilhelm Krelle; Wulf Gaertner; Constantino Lluch; Michael McCarthy; Phlips (1974, chs. 6, 7, and 10); Ahmad El-Safty (1976a,b); Peter Hammond; Nico Klijn. Empirical investigations based on various specifications of habit formation include the author and Terence Wales; Wales; Phlips (1972); Murray Brown and Dale Heien; Taylor and Daniel Weiserbs; Richard Boyce; Marilyn Manser.

sumption in the previous period), but a specification which assigns declining geometric weights to all past consumption is equally tractable. With the one-period lag specification, we denote the household's preference ordering in period t by $R(Q_{t-1})$, where Q_t is the consumption vector for period t. The statement $Q_t^a R(\bar{Q}_{t-1}) Q_t^b$ means that the household finds Q_t^a at least as good as Q_t^b given the consumption history $\bar{Q}_{t-1}$. The "short-run demand functions" are denoted by $Q_t = h(P_t, \mu_t, Q_{t-1})$, where P_t is the vector of prices p_i in period t and μ_t is total expenditure in period t. The "long-run demand functions" $Q = H(P,\mu)$ are defined to be the steady-state solution to the short-run demand functions: $H(P,\mu) = h[P,\mu, H(P,\mu)]$.

For example, the familiar Klein-Rubin linear expenditure system

$$h^i(P,\mu) = b_i - \frac{a_i}{p_i}\sum_{k=1}^{n} p_k b_k + \frac{a_i}{p_i}\mu$$
$$\Sigma a_k = 1$$

is generated by the direct utility function

$$U(Q) = \sum_{k=1}^{n} a_k \log(q_k - b_k),$$
$$a_i > 0, (q_i - b_i) > 0, \Sigma a_k = 1$$

For this demand system a particularly simple specification of habit formation is one in which the b's depend linearly on consumption in the previous period. Under this specification, the utility function becomes

$$U(Q_t; Q_{t-1}) = \sum_{k=1}^{n} a_k \log(q_{kt} - b_k^* - \beta_k q_{kt-1}),$$
$$a_i > 0, (q_{it} - b_i^* - \beta_i q_{it-1}) > 0,$$
$$\Sigma a_k = 1$$

and the corresponding short-run demand functions are given by

$$h^i(P_t,\mu_t,Q_{t-1}) = b_i^* + \beta_i q_{it-1} - \frac{a_i}{p_{it}}\sum_{k=1}^{n} \cdot p_{kt}(b_k^* + \beta_k q_{kt-1}) + \frac{a_i}{p_{it}}\mu_t$$

Interdependent preferences—that is, preferences which depend on the consumption decisions of others—have received little attention from economists.[3] The intrinsic difficulty is that everything depends on everything else: A's tastes depends on B's consumption, and vice versa, so their equilibrium consumption patterns must be determined simultaneously. But this difficulty can be avoided by following the tack taken in the habit formation literature: A's preferences are assumed to depend on B's *past* consumption, and vice versa. This specification seems to capture the essence of interdependent preferences, although the short-run demand functions are not simultaneously interdependent. In the long run, everyone's preferences depend on everyone else's consumption, so we are led to consider per capita consumption patterns for the entire society as well as individual demand behavior. Although habit formation and interdependent preferences imply different patterns of individual behavior, the per capita demand functions implied by some specifications of interdependent preferences are indistinguishable from those implied by certain specifications of habit formation.

II. Taste Change and Welfare Analysis

Welfare comparisons require a different interpretation of preferences than demand analysis. In demand analysis the objects of choice are vectors of private decision variables Q, and preferences over them depend on a vector of predetermined "state variables" Z; we call such a preference ordering

[3]The post-Veblen literature on interdependent preferences begins with James Duesenberry's well-known book; Robert Clower points out some difficulties in Duesenberry's formulation. Harvey Leibenstein (1950) and S. J. Prais and Houthakker (ch. 2) discuss interdependent preferences, but the subject appears to have been dormant from the mid-1950's until the early 1970's when it was taken up by Krelle; Gaertner; the author (1976a); Hiroaki Hayakawa and Yiannis Venieris. The exposition in the text follows the author (1976a). Models of taste formation also play a major role in the analysis of fertility. See Richard Easterlin (1973, 1976); Easterlin, the author, and Michael Wachter; Leibenstein (1975, 1976).

"conditional." For example, with simultaneous interdependent preferences, the private decision variables are goods consumed by the individual and the state variables are everyone else's current consumption. We denote the individual's conditional preference ordering by $R(Z)$ and interpret the statement $Q^a R(\bar{Z}) Q^b$ to mean that Q^a is at least as good as Q^b when other people's consumption is given by $\bar{Z}$. Since the individual takes other people's consumption as fixed, demand analysis need never ask how an individual would choose between alternatives which differ with respect to other people's consumption; hence, conditional preferences are an adequate foundation for demand analysis.

Welfare analysis must compare the individual's well-being in alternative situations which differ with respect to the state variables as well as the private decision variables. For example, we might ask whether the individual is better off in the status quo or an alternative situation in which his consumption is 10 percent higher while everyone else's is 20 percent higher. This comparison cannot be based on conditional preferences but requires a framework in which the objects of choice include other people's consumption as well as his own. In general, welfare analysis requires us to redefine the objects of choice to include not only the private decision variables but also the state variables which, from the standpoint of conditional preferences, are predetermined. We call an ordering over such an augmented set of alternatives an "unconditional preference ordering," and denote it by R: the statement $(Q^a, Z^a) R (Q^b, Z^b)$ means that the individual finds (Q^a, Z^a) at least as good as (Q^b, Z^b). The additional information contained in the unconditional preference ordering is irrelevant to demand analysis, but for welfare evaluation it is indispensable.

The state variables may be the individual's own past consumption (habit formation) or the consumption of others (interdependent preferences). They may be environmental variables (for example, climatic conditions or pollution), health states, or goods or services provided by the government (for example, highways, schools, recreational facilities). Or they may be socioeconomic variables or demographic variables, such as the number, age, and sex of the children in a family. In all of these cases, conditional preferences provide a foundation for analyzing the effect of the state variables on market behavior, but welfare analysis requires unconditional preferences. For example, the effect on an individual's welfare of a 20 percent increase in everyone else's consumption cannot be inferred from its effect on his consumption pattern.

Unconditional preferences are necessary for welfare analysis, yet they cannot be inferred from market behavior or any other conditional choices. How, then, might they be discovered? In some cases, we can observe the individual's unconditional choices, although these are not market choices. For example, with interdependent preferences we might infer an individual's unconditional preferences from his willingness to support alternative redistributional tax and transfer schemes. Otherwise, the problem of identifying unconditional preferences has much in common with that of making interpersonal comparisons of well-being.[4]

[4]On welfare evaluation with changing tastes, see John Harsanyi; Franklin Fisher and Karl Shell; von Weizsäcker; the author (1976b). The distinction between conditional and unconditional preferences is developed in my 1976a and 1977 articles. On interpersonal comparisons, see Amartya Sen (pp. 13–15) and Kenneth Arrow (pp. 224 25).

REFERENCES

K. J. Arrow, "Extended Sympathy and the Possibility of Social Choice," *Amer. Econ. Rev. Proc.*, Feb. 1977, *67*, 219–25.

R. Boyce, "Estimation of Dynamic Gorman Polar Form Utility Functions," *Annals Econ. Soc. Measure.*, Winter 1975, *4*, 103 16.

M. Brown and D. Heien, "The S-Branch Utility Tree: A Generalization of the Linear Expenditure System," *Econometrica*, July 1972, *40*, 737–47.

R. W. Clower, "Professor Duesenberry and Traditional Theory," *Rev. Econ. Stud.*,

No. 3, 1952, *19*, 165–78.

James S. Duesenberry, *Income, Saving, and the Theory of Consumer Behavior*, Cambridge 1949.

R. A. Easterlin, "Relative Economic Status and the American Fertility Swing," in Eleanor B. Sheldon, ed., *Family Economic Behavior: Problems and Prospects*, Philadelphia 1973, 170–223.

———, "The Conflict Between Aspirations and Resources," *Pop. Develop. Rev.*, Sept./Dec. 1976, *2*, 417–25.

———, **R. A. Pollak, and M. L. Wachter**, "Toward a More General Economic Model of Fertility Determination: Endogenous Preferences and Natural Fertility," in *Population and Economic Change in Less Developed Countries*, Universities-Nat. Bur. Econ. Res. conference series, forthcoming.

A. E. El-Safty, (1976a) "Adaptive Behavior, Demand and Preferences," *J. Econ. Theory*, Oct. 1976, *13*, 298–318.

———, (1976b) "Adaptive Behavior and the Existence of Weizsäcker's Long-Run Indifference Curves," *J. Econ. Theory*, Oct. 1976, *13*, 319–28.

F. M. Fisher and K. Shell, "Taste and Quality Change in the Pure Theory of the True Cost-of-Living Index," in J. N. Wolfe, ed., *Value, Capital and Growth: Papers in Honour of Sir John Hicks*, Edinburgh 1968, 97–139.

Milton Friedman, *Price Theory: A Provisional Text*, Chicago 1962.

Wulf Gaertner, "A Dynamic Model of Interdependent Consumer Behavior," *Z. Nationalökon.*, No. 3–4, 1974, *34*, 327–44.

John Kenneth Galbraith, *The Affluent Society*, Cambridge 1958.

———, "Economics as a System of Belief," *Amer. Econ. Rev. Proc.*, May 1970, *60*, 469–78.

H. Gintis, "Welfare Criteria with Endogenous Preferences: The Economics of Education," *Int. Econ. Rev.*, June 1974, *15*, 415–30.

W. M. Gorman, "Tastes, Habits, and Choices," *Int. Econ. Rev.*, June 1967, *8*, 218–22.

P. J. Hammond, "Endogenous Tastes and Stable Long-Run Choice," *J. Econ. Theory*, Oct. 1976, *13*, 329–40.

J. S. Harsanyi, "Welfare Economics of Variable Tastes," *Rev. Econ. Stud.*, No. 3, 1954, *21*, 204–13.

H. Hayakawa and Y. Venieris, "Consumer Interdependence via Reference Groups," *J. Polit. Econ.*, June 1977, *85*, 599–615.

Hendrik S. Houthakker and Lester D. Taylor, *Consumer Demand in the United States: Analyses and Projections*, Cambridge, Mass. 1966; 2d ed., 1970.

N. Klijn, "Expenditure, Savings, and Habit Formation: A Comment," *Int. Econ. Rev.*, Oct. 1977, *18*, 771–78.

W. Krelle, "Dynamics of the Utility Function," in John R. Hicks and Warren Weber, eds., *Carl Menger and the Austrian School of Economics*, New York 1973, 92–128.

H. Leibenstein, "Bandwagon, Snob, and Veblen Effects in the Theory of Consumer Demand," *Quart. J. Econ.*, May 1950, *64*, 183–207.

———, "The Economic Theory of Fertility Decline," *Quart. J. Econ.*, Feb. 1975, *89*, 1–31.

———, "The Problem of Characterizing Aspirations," *Pop. Develop. Rev.*, Sept./Dec. 1976, *2*, 427–31.

C. Lluch, "Expenditure, Savings and Habit Formation," *Int. Econ. Rev.*, Oct. 1974, *15*, 786–97.

M. D. McCarthy, "On the Stability of Dynamic Demand Functions," *Int. Econ. Rev.*, Feb. 1974, *15*, 256–59.

M. E. Manser, "Elasticities of Demand for Food: An Analysis Using Non-Additive Utility Functions Allowing for Habit Formation," *Southern Econ. J.*, July 1976, *43*, 879–91.

M. H. Peston, "Changing Utility Functions," in Martin Shubik, ed., *Essays in Mathematical Economics in Honor of Oskar Morgenstern*, Princeton 1967, 233–36.

Louis Phlips, "A Dynamic Version of the Linear Expenditure Model," *Rev. Econ. Statist.*, Nov. 1972, *54*, 450–58.

———, *Applied Consumption Analysis*, Amsterdam 1974.

R. A. Pollak, "Habit Formation and Dynamic Demand Functions," *J. Polit. Econ.*, July/Aug. 1970, *78*, 745–63.

———, (1976a) "Interdependent Preferences," *Amer. Econ. Rev.*, June 1976, *66*, 309-20.

———, (1976b) "Habit Formation and Long-Run Utility Functions," *J. Econ. Theory*, Oct. 1976, *13*, 272-97.

———, "Price Dependent Preferences," *Amer. Econ. Rev.*, Mar. 1977, *67*, 64-75.

——— **and T. J. Wales,** "Estimation of the Linear Expenditure System," *Econometrica*, Oct. 1969, *37*, 611-28.

S. J. Prais and Hendrik S. Houthakker, *The Analysis of Family Budgets*, Cambridge 1955.

Richard Schmalensee, *The Economics of Advertising*, Amsterdam 1972.

Amartya K. Sen, *On Economic Inequality*, London 1973.

G. J. Stigler and G. S. Becker, "De Gustibus Non Est Disputandum," *Amer. Econ. Rev.*, Mar. 1977, *67*, 76 90.

Richard Stone, "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand," *Econ. J.*, Sept. 1954, *64*, 511-27.

———, (1964a) "The Changing Pattern of Consumption," in *Problems of Economic Dynamics and Planning*, Warsaw 1964.

———, (1964b) "British Economic Balances in 1970: A Trial Run on Rocket," in P. Hart et al., eds., *Econometric Analysis for National Planning: 16th Symposium of the Colston Society*, London 1964.

L. D. Taylor and D. Weiserbs, "On the Estimation of Dynamic Demand Functions," *Rev. Econ. Statist.*, Nov. 1972, *54*, 459 65.

C. C. von Weizsäcker, "Notes on Endogenous Change of Tastes," *J. Econ. Theory*, Dec. 1971, *3*, 345-72.

T. J. Wales, "A Generalized Linear Expenditure Model of the Demand for Non-Durable Goods in Canada," *Can. J. Econ.*, Nov. 1971, *4*, 471-84.

Stochastic Properties of Changing Preferences

By EDGAR A. PESSEMIER*

This discussion of the dynamics of preference and choice is based principally on the research work of psychologists and marketing scholars. Psychologists have advanced a number of useful theories of individual perceptual and affective processes, and tested these theories in a variety of experimental settings. On the other hand, marketing scholars have been interested in formulating efficient marketing strategies; marketing efforts that must be directed to groups of individuals. Therefore, aggregate measures such as store traffic counts or product sales are the principal subjects of their investigations.

It is not surprising to find psychologists interested in the applied implications of their theories or to find marketing investigators concerned about theories of individual behavior which can provide additional insight into aggregate behavior. The work of J. Douglas Carroll and Paul Green illustrates productive collaboration based on these common interests. Although Kevin Lancaster's work has extended the deterministic theory of consumer demand, it is surprising to note the minor role played by economists in expanding knowledge about individual choice behavior or about strategy formulation by individual firms. In large part, this condition is due to the absence of a strong empirical/experimental tradition in economics and to the economists's preoccupation with aggregate data bearing on very broad policy questions.

I. Some Basic Concepts

The easiest way to look at individual preference and choice is to start with three basic concepts: evoked sets; preceptions; preferences.

*Purdue University.

A. *The Evoked Set*

In any particular choice situation, an evoked set includes those objects of choice which are considered by an individual prior to actively engaging in a choice decision. It is useful to think of this set as the "planning stage" object set. In some respects, it is analogous to Henri Theil's planning stage decision variable set. Brian Campbell has discussed determinants of the evoked set and the concept has been applied in a number of marketing research studies, usually in the form of various measures of awareness. The latter measure may concern either objects or the properties of objects.

The evoked set usually varies from individual to individual, and from one occasion to another. Also, the evoked set typically contains only a small fraction of the objects found in the associated market. Finally, G. A. Miller notes that an evoked set rarely contains more than ten objects.

B. *Perceptions*

An individual's preceptions of an object defines his or her beliefs or judgments about the properties of the object. These beliefs may concern simple objectively measurable properties such as weight or brightness or an abstract composite, such as similarity. In a market context, subjective properties such as sportiness, safety, reliability, and value tend to be of greater interest. These latter characteristics often summarize how a group of related design attributes are internalized by an individual. Perceptions of objects' determinant attributes occupy a central role in various unfolding or joint space models of individual choice behavior (see Green and Vithala Rao; Green and Yorham Wind; the author, 1977). These latter attributes influence preference *and* vary noticeably from object to object.

Generally, beliefs about objects change slowly and are more widely shared than preferences for objects. Nevertheless, perceptions of an object vary from individual to individual and from occasion to occasion. The literature in psychophysics reports on a vast body of experimental evidence concerning the variability of individual judgments about objects (see R. Darrell Bock and Lyle Jones; Clyde Coombs; Warren Torgerson). Among the earliest and best known theoretical treatment of this variability is L. L. Thurstone's Law of Comparative Judgment and the related concepts of discriminal dispersion.

C. *Preference*

An individual's preferences for objects define each object's relative appeal. An individual's probability of choosing an object on the next choice occasion is strongly associated with these predispositions. A large empirical literature has been built on the early theoretical work by R. A. Bradley, M. E. Terry and R. Duncan Luce (see Bock and Jones). More recently the focus has been on the functional relationship between an individual's perceptions of the objects' determinant attributes and affective judgments. The latter judgments flow from an individual's evaluative process operating in the context of his or her personal needs and resources (see the author, 1977, 1974).

Manifest preference has been used to predict choice in a wide range of market related choice situations. Although the measurement methods and theoretical models have varied greatly, many models employ the deviation of an object from an ideal level of the attribute as the determinant of preference (see Green and Rao; Green and Wind; the author and William Wilkie). Some controversy has surrounded the question of attainable predictive accuracy, but little support can be found for treating choice as a fully rational deterministic process. Furthermore, the relative nature of preference and choice has been convincingly demonstrated. The number and character of the choice objects being considered strongly influences manifest preference and choice. Carroll; David Bell, Ralph Kenney, and John D.C. Little; Luce; and the author have discussed this subject from somewhat different points of view.

Finally, preferences vary from individual to individual and from one occasion to another. This latter variability may be due to changing content of the evoked set, changing perceptions of objects in the evoked set, changing affective weights assigned to the attributes of objects, and to changes in the individual's needs and resources.

D. *Summary*

The above discussion and the weight of current evidence support a dynamic view of preference and choice. In general, it is a hierarchical view that flows from awareness, comprehension, and evaluation to choice. Changes in the first three elements tend to induce changes in choice. Furthermore, this perspective is strongly stochastic, suggesting that individual choice on any single choice occasion cannot be accurately predicted under steady-state conditions. The perturbing influences of the "neglected variables" related to a specific purchase are often very influential and difficult to ignore, particularly in choice situations where object preferences are weakly structured. Unfortunately, it is nearly impossible to effectively measure and incorporate these situational variables in a formal model. Even if these variables could be disregarded, selection on the next purchase occasion can seldom be predicted with satisfying accuracy. Aggregate predictions across individuals or occasions, however, can accurately forecast the share of choices received by various objects (see the author et al.).

Finally, Theil's notion about the planning stage and the implementation stage may be adapted to fit the above choice context. In this case, tradeoffs can be made between the expected utility from the prior evaluation stage and flexibility of decision at the time of choice (as influenced by the

neglected variables). Although more will be said at a later point about both the sources and value of the variable nature of individual choice behavior, at this point it is worth emphasizing that even if the dynamic and stochastic elements could be eliminated, which they cannot, it seems quite clear that it would be unwise to do so (see the author, 1975).

II. Models of Preference and Choice

As noted earlier, a number of formal spacial models of individual choice behavior have been developed that incorporate differences in the affective processes of individuals and the different properties of choice objects (see Carroll; the author and Wilkie; Amos Tversky). Marketing scientists have converted these approaches into useful predictive models that jointly represent the effects of each market segment's particular needs and each firm's product and communications strategy. By making proper allowance for the observed heterogeneity of demand within a product category, these models can help a firm respond with a uniquely suitable competitive set of products and associated marketing programs.

A. *Basic Structure*

The structure of the above models calls for locating objects (products) in a perceptual space spanned by a modest number of relatively independent components that can be computed from the perceived positions of products on their underlying determinant attributes. These latter attributes are closely associated with real design variables and may be manipulated to influence both product perceptions and cost. An alternative way to analyze the independent composite attributes that span the perceptual space is to specify these characteristics, for example, sportiness and operating economy, in a separate study. Then each characteristic can be judged or treated as a dependent variable in a linear model design to explain the level of each product on each characteristic in terms of the constituent attributes. However a reduced space representation of choice objects may be developed, it becomes a joint space when product preference measures are used to assign a (most preferred) location to each (preference homogeneous) market segment. It is convenient to think of this joint space as a market map since it contains all the basic information needed to analyze the strategic actions available to firms that compete in relatively mature markets.

B. *Changing Structure*

The market-map model is less useful in cases where customer product knowledge is very weak or just developing around an innovative product or product category. In the latter case, awareness of both the product(s) and its (their) determinant attributes will increase as the adoption increases. When a product or product class is very new, individuals who indicate the product is in their evoked set or who are willing to judge the properties of the product may do so less from real knowledge than from a desire to supply the investigator with data. In other words, early in the adoption process the principal questions concern the initial responses at low knowledge levels, the attainable levels of knowledge, the rate and type of knowledge that is likely to be transmitted, and the effectiveness of various efforts to influence the rates of change and the attainable levels of knowledge. Under these conditions, it would be useful to have a market-mapping method which analyzes the changing awareness of individuals about products and their determinant attributes. One possibility currently being investigated involves mapping contingency data (yes-no judgments about product attributes) for each market segment at various points during the adoption process. Following Paul Ries, this approach builds on the theoretical contributions of J.P. Benzécri. Other models and measurements can be suggested. The main point is that formally introducing the effects of time dependent changes in the perceptual and/or affective process can greatly enhance the value of market-mapping methods.

C. *Summary*

The rapid development of models of preference during the past several decades has greatly increased understanding of individual preference and choice behavior and greatly enriched the analysts understanding of market behavior. The associated analytical methods and planning algorithms are being actively evaluated by applied researchers and tested in real market conditions. The diagnostic and predictive quality of these models will improve as applied researchers learn which sampling, measurement, and analytical techniques are required in various situations.

At a more fundamental level, experience and further development can be expected to yield relatively accurate predictions of aggregate choice and sales behavior. In this process, analysts should not lose sight of the inherent and substantial dynamic and stochastic components of individual market choices. Careful attention should also be given to understanding the heterogeneity of preferences within each population under study.

Since uncertainty, change, and uniqueness are important aspects of buyer and seller behavior, it is useful to explore several factors which may encourage these modes of behavior. The following discussion emphasizes potential biological and psychological explanations.

III. Uncertainty, Change, and Uniqueness

Market choices involve risk for the decision maker. This risk can be reduced by collecting additional information but the cost of information and the amount of residual uncertainty are usually high. These conditions encourage the use of behavioral information strategies as well as analytical search strategies. As M. J. Klingsporn and Jacques Monod have argued, behavioral strategies predominate in nature. They lie behind the dynamics of adoption for all living entities. In the face of an uncertain future, varied responses are essential. Out of this variety, the new environment identifies viable behavior.

A. *Market Risks and Responses*

Conditions in the market place parallel those encountered by living systems. Sellers cannot fully anticipate how products will be received by resellers and consumers. Varied product forms are frequently offered to learn about the demands of various groups. As a further protection against market uncertainties, a firm may seek a larger, more diverse clientele to reduce the cost of failure with a single group of buyers. In part these actions are a game against competitors and the market environment. In another sense they are an efficient information strategy for examining the effectiveness of actions in regions adjacent to those that have been explored by current experience.

Individual customers face information problems that are not unlike those faced by sellers. They are never sure that any purchase will be the best choice. If a satisfactory product is repeatedly purchased, the individual slowly loses the capacity to judge the product's relative desirability. The buyer, like the seller, needs a behavioral strategy as well as an analytical search strategy. A degree of randomness or experimentation is called for. The most suitable level is determined by the rate at which habituated behavior becomes dysfunctional. This latter rate depends on how fast and to what extent market offerings and individual needs shift.

Varied, possibly random, responses by buyers and sellers impose real risks but they are a strong defense against progressive obsolescence. Ill-understood, volatile, risky conditions call for more intense, carefully considered exploratory behavior and faster response time. Less exploration and slower response times are appropriate for more stable, better understood situations.

B. *Market Diversity*

Investigators in many fields have observed that increased diversity implies greater energy requirements (see Kenneth Watt, 1973, 1974). In an economic context, increased specialization and scale can

reduce energy consumption. In a period when energy conservation is becoming a serious national concern, there is likely to be a growing impulse to limit the varied responses of buyers and sellers. For the reasons noted above, such a policy will increase the potential obsolescence of the smaller, more stable collections of choice objects and more limited purchase repertoires of buyers.

Restricting the number of choice objects or the speed with which they are replaced with new versions has another unfavorable effect. Individuals need distinct collections of personal property to establish their individuality and social identity. Furthermore, Howard Fromkin observed that most people enjoy variety in possessions as much as they do in the arts and personal life styles. The individual's need for variety accounts for some of the observed contrasts in the attributes of objects chosen on adjacent purchase occasions.

C. *Summary*

Individual preference behavior is highly varied and choice behavior cannot be easily predicted on any given purchase occasion. These observations do not flow from irrational behavior, faulty measurement or imperfect model building. A varied choice repertoire can be attributed largely to the adaptive characteristics and social needs of the species. Instead of ignoring or demeaning these inconvenient facts of life, their economic effects should be examined with care.

REFERENCES

D. Bell, R. Keeney, and J. D. C. Little, "A Market Share Theorem," *J. Marketing Res.*, May 1975, *12*, 136–41.

J. P. Benzécri, *L'analyse des donnees*; Vol. 2 *L'analyse des correspondences*, Paris 1973.

R. Darrell Bock and Lyle V. Jones, *The Measurement and Prediction of Judgment and Choice*, San Francisco 1968.

B. M. Campbell, "The Existence and Determinants of Evoked Sets in Brand Choice Behavior," unpublished doctoral dissertation, Columbia Univ. 1969.

J. Carroll, "Individual Differences and Multidimensional Scaling," in Roger Shepard et al., eds., *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, Vol. 1, New York 1972, 105–55.

Clyde H. Coombs, *A Theory of Data*, New York 1964.

H. L. Fromkin, "The Psychology of Uniqueness; Avoidance of Similarity and Seeking of Differences," Krannert Grad. Sch. Manage., inst. paper no. 438, Purdue Univ. 1974.

Paul E. Green and J. Douglas Carroll, *Mathematical Tools for Applied Multi-variate Analysis*, New York 1976.

——— **and Vithala Rao,** *Applied Multidimensional Scaling*, New York 1972.

——— **and Yorham Wind,** *Multiattribute Decisions in Marketing*, Hinsdale 1973.

M. J. Klingsporn, "The Significance of Variability," *Behav. Sci.*, Nov. 1973, *18*, 441–47.

Kevin Lancaster, *Consumer Demand*, New York 1971.

R. Duncan Luce, *Individual Choice Behavior*, New York 1950.

G. A. Miller, "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychol. Rev.*, 1956, *63*, 81–97.

Jacques Monod, *Chance and Necessity*, New York 1971.

Edgar A. Pessemier, *Product Management: Strategy and Organization*, New York 1977.

———, "Market Darwinism, Choice Theory and Marketing Models," *Amer. Marketing: Fall Conf. Proc.*, 1975, 27–30.

——— **and W. Wilkie,** "Multi-Attribute Choice Theory—a Review and Analysis," in G. David Hughes and Michael Ray, eds., *Buyer/Consumer Information Processing*, Chapel Hill 1974, 288–330.

——— **et al.,** "Using Laboratory Brand Preferences to Predict Consumer Brand Purchases," *Manage. Sci.*, Feb. 1971, *17*, B371–85.

P. N. Ries, "Joint Space Analysis of Contingency Data," unpublished paper, Procter and Gamble Co., Cincinnati, Oct. 1974.

H. Theil, "A Theory of Rational Random Behavior," *J. Amer. Statist. Assn.*, June 1974, *69*, 310–14.

L. L. Thurstone, *The Measurement and Prediction of Value*, Chicago 1959.

Warren S. Torgerson, *Theory and Methods of Scaling*, New York 1958.

A. Tversky, "Choice by Elimination," *J. Math. Psychol.*, Nov. 1972, *9*, 341–67.

K. E. F. Watt, "Men's Efficient Rush Towards Deadly Dullness," *Nature*, Feb. 1974, 74–82.

——— "A Movable (Disappearing) Feast," *Saturday Rev.*, Feb. 1973, 55–56.

On the Study of Taste Changing Policies

By T. A. MARSCHAK*

To enter the field of taste changes one ought to find danger exhilarating. The perils are extreme. First, the very ground threatens to fall away at one's feet: the economist, as policy adviser, is supposed to seek efficiency, but whether a given policy is efficient depends upon the preferences of those affected, and those preferences may depend in turn on policy. Second, if one continues to believe that even in a world of changeable tastes the foundation for policy and prediction has to be a theory of individual rational choice, then one risks turning Economic Man into a complex monster of calculated schizophrenia, who chooses or manipulates future mutations of himself. Third, and most alarming of all, one risks discovering that true progress in this field means entering long-forbidden territory; exploring the structure of human contentment from the inside; searching beyond the hints about what people want that are given by the old familiar economic observables (prices, incomes, and quantities demanded); and becoming at last full and active partners with "behavioral scientists," unrigorous as they may be, shaping their work as well as learning from it.

Love of danger aside, motivation for the field mounts and it has, indeed, grown surprisingly in a few years. At least three major policy issues have to do with taste changes. First, if there are indeed important nonrenewable resources, then it becomes crucial to determine how present consumption of the resource, and policies which discourage it, affect future tastes. Second, if one feels that advertising, education, and mass media are all taste altering, and their use of resources a policy matter, then assessing alternative policies (one policy is leaving everything alone) means studying how tastes change.

A third issue is difficult and touchy. That is whether, in a developed economy, a massive shift by many people in the way they choose to use their time, a shift towards "leisure" and away from the exchange of time for purchased goods, means a gain or a loss in social welfare. This issue seems to lie at the core of popular debates about "lowered aspirations" or (a recent slogan) "voluntary simplicity" or the contention that "small is beautiful." Some claim that such a shift is actually under way in American society and that it would in any case be some sort of national salvation. Others claim there are no serious signs of such a shift and that it would in any case be a national (even worldwide) disaster. It is often a loose emotional debate, lacking a middle ground.

Not all economists have ignored the issue. It appears prominently in the so-called growth/no-growth controversy, to which the latest contribution is E. J. Mishan's impassioned new book: growth as traditionally measured cannot, we are told, achieve the good life; a change in the way individuals use their time may achieve it. Staffan Linder argues that since time is required to consume the products of ever more productive working hours, an individual's "spare" time—neither working time nor goods-consumption time—may be squeezed away and contentment diminished; a taste shift toward spare time and away from goods might resolve the dilemma. Tibor Scitovsky's recent innovative book uses findings from motivational psychology to start categorizing basic satisfactions which people derive from goods and activities, and to start exploring why some societies appear better at transforming time into contentment than other societies.

I shall focus on the last of the three complex slippery policy issues to explore how the taste change literature tells us to study it. Suppose we take seriously the most primitive family of static models which permits one to ask precise questions about the

*University of California-Berkeley.

welfare effect of policies that cause many people to shift from "goods" to leisure, including policies that appear to change tastes directly.

I. Leisure vs. Goods Policies in a One-Person Economy

We begin with the simplest model of all: a one-person economy—Crusoe—with a single produced commodity (called goods). The twenty-four hours in a day are to be divided between leisure and goods production. There is a production function h; t hours of time not devoted to leisure yields $h(t)$ units of goods. We are concerned with Crusoe's preferences over alternative daily goods-leisure combinations and with policies that might affect these preferences. There are only two periods and a policy is applied in period 1; if it affects Crusoe's tastes at all, then it affects them in period 2. Assume that 1) Crusoe is the traditionally careful utility maximizer and 2) while in period 1 he may be indifferent to his fate in period 2, yet he always *knows* what his period 2 preferences will be.

Then there are three Crusoes - A, B, and C - whose views might be considered in recommending a policy: A is Crusoe in period 1; B is Crusoe as he would be in period 2 if the policy were adopted, given the choices A would then make in period 1; C is Crusoe as he would be in period 2 if the policy were not adopted. One finds in the literature various views as to who needs to "accept" the policy– to find himself no worse off under the policy than without it – before it can be judged an improvement. One view is that A and B should accept (see Burton Weisbrod, reinterpreted in the present context). A second view (see Menahem Yaari, reinterpreted) is that the utility of B and C should be meaningfully comparable, A should accept the policy, and B's utility under the policy should be higher than C's without it. A third view (see C. Christian von Weizsäcker), confined to "endogenous" taste changes, is that B's acceptance alone matters provided that under the policy he makes a steady-state choice– a choice he would repeat in a third period, if there were one. According to a fourth view (see von Weizsäcker, generalized), B should find his best choice after the policy to be better than A's choice was without it.

A truly cautious judge would, in the spirit of ordinalist welfare economics, reject all these viewpoints, since some of them require "intrapersonal" comparisons and none require acceptance of the policy by C, the person whom in effect the policy deprives of existence. The most cautious view requires acceptance of the policy by A, B, and C. However it is obvious—and will be illustrated—that such unanimity is impossible for truly taste altering policies which preserve (or diminish) the set of choices attainable in period 2.

Let L_i denote Crusoe's leisure in period i (in hours) and G_i goods consumed in period i. Consider several preference structures and policies: (a) Suppose first that utility in period i is $u(L_i, G_i, \lambda_i)$, increasing in L_i, G_i; λ_i is a parameter and increasing it means raising everywhere the marginal rate of substitution of leisure for goods. More precisely, if $\lambda'' > \lambda'$, then whenever $u(L^*, G^*, \lambda') = u(L, G, \lambda')$ and $u(L^{**}, G^*, \lambda'') = u(L, G, \lambda'')$, we have $L^{**} < L^*$. A proposed policy (advertising) imposes in period 2, $\lambda_2 > \lambda_1$. Without the policy, $\lambda_2 = \lambda_1$. Then A (period 1 Crusoe) accepts the policy since he is indifferent to the future. And B (period 2 Crusoe under the policy) accepts it, since his best attainable utility (under λ_2) is higher than $u[\hat{L}_2(\lambda_1), h(24 - \hat{L}_2(\lambda_1)), \lambda_2]$, where $\hat{L}_2(\lambda)$ denotes utility-maximizing period 2 leisure for the parameter value λ. However, normally C (the period 2 Crusoe who would have existed without the policy) rejects it, since normally

$$(1)\quad u[\hat{L}_2(\lambda_2), h(24 - \hat{L}_2(\lambda_2)), \lambda_1] < u[\hat{L}_2(\lambda_1), h(24 - \hat{L}_2(\lambda_1)), \lambda_1]$$

(b) Now suppose we learn something new about Crusoe's preferences. They have all the properties of case (a), but we now know in addition that the period i utility function is of the Gary Becker household-technology sort. It has the form $\bar{u}(f(L_i, \lambda_i), g(G_i))$, where $\bar{u}$, f, and g are each increasing

in all arguments; f and g are production functions yielding a "leisure commodity" (requiring only time as an input) and a "goods commodity" (requiring only goods). The proposed policy is as before, but we now know that the parameter it shifts affects "leisure productivity" only (perhaps the policy supplies facilities permitting better use of leisure). Increasing leisure productivity means that every previously available leisure-commodity/goods-commodity pair can now be improved upon (with regard to $\bar{u}$). Hence not only A and B but C as well now accept the policy. It is, in fact, a policy not affecting Crusoe's true tastes at all.

The inequality (1) still holds, but no longer implies a rejection of the policy by C. If the utility function has the household-technology form $\bar{u}$, then the Crusoes described in case (a) with λ_1 *or* λ_2 are merely alternative *codings* of the household-technology Crusoe; replacing λ_1 by λ_2 does not change this person's true tastes.

The true tastes, as stressed in taste change writings of Peter Hammond, are given by a set of choosable objects and a choice function which selects elements from any subset of the set. Suppose S is a set of pairs (X, Y) (leisure-commodity/goods-commodity pairs) given to the household-technology Crusoe. Consider the set $m_\lambda(S)$ of all pairs (r,s) (leisure time/goods pairs) where for some (X,Y) in S, $X = f(r,\lambda)$, $Y = g(s)$. Suppose instead of describing Crusoe as someone who chooses out of S the pair $(\bar{X},\bar{Y})$ (assume it unique) for which $\bar{u}(X,Y)$ is maximal, we describe him as someone who first computes the set $m_\lambda(S)$ and then chooses out of that set the pair $(\bar{r},\bar{s})$ for which $\bar{u}(r,s,\lambda) = u(f(r,\lambda),g(s))$ is maximal. *Whatever* λ *may be*, we have described the same person, since $\bar{X} = f(\bar{r},\lambda)$, $\bar{Y} = g(\bar{s})$. What first appears to be a taste altering policy turns out—once we *add* to our knowledge of preferences the fact that they have the household-technology structure—to be merely a recoding of a person whose tastes are unchanged.

(c) Suppose Crusoe does care about the future. Period 1 utility is $u_1(L_1,G_1,L_2,G_2)$, increasing in all arguments. Period 2 utility is $u_2(L_2, G_2, \lambda)$, where an increase in λ again increases the marginal rate of substitution of leisure for goods. The imposed policy lets Crusoe increase his parameter from its status quo value of λ to $\lambda + K$, by using $\theta(K)$ hours of period 1 time ($\theta(K) + L_1 \leq 24$). (The policy provides, say, time-consuming educational opportunities which shift next period's preferences towards leisure.) To find period 1 Crusoe's utility under the policy, we have to find period 2 Crusoe's best choice of L_2, say $\hat{L}_2(K)$, given any period 1 choice of K. Then A's utility is $u_1[\bar{L}_1,h(24 - \bar{L}_1 - \theta(\bar{K})),\hat{L}_2(\bar{K}),h(24 - \hat{L}_2(\bar{K}))]$, where $\bar{L}_1$, $\bar{K}$ are maximizers. This is not less than A's utility without the policy (when $K = 0$). Crusoe B, with tastes given by $\lambda + \bar{K}$, of course accepts the policy (prefers $[\hat{L}_2(\bar{K}),h(24 - \hat{L}_2(\bar{K}))]$ to the (L_2,G_2) pair which would be chosen when $K = 0$). Crusoe C, with tastes given by λ, normally rejects it.

(d) Again we further restrict (c) to achieve a household-technology version. Period 1 utility is $u_1[f(L_1,\lambda),g(G_1),u_2(f(L_2,\lambda),g(G_2))]$ (increasing in all three arguments). Period 2 utility is $u_2[f(L_2,\lambda),g(G_2)]$. The same policy as before is now accepted by A, B, and C since it is now not a taste altering policy at all.

(e) Consider an "endogenous change" example. Period 1's utility is $u_1(L_1,L_2,G_1,G_2)$ and period 2's is $u_2(L_1,L_2,G_2)$, increasing in L_2,G_2. Increasing L_1 increases everywhere the marginal rate of substitution of period 2 leisure for period 2 goods. A proposed policy provides a temporary "subsidy" of S hours of period 1 leisure: the goods previously produced with R hours of labor ($R > S$) are now produced with $R - S$ hours. Crusoes A and B accept the policy (A's utility is determined analogously to case (c)), while C, whose tastes are those determined by the presubsidy period 1 choice of leisure, normally rejects it.

Here one can—more reasonably than in exogenous change examples—require that B's utility and C's be comparable, and can judge the policy an improvement if B's utility under it is higher than C's without it.

This means that Crusoe is able to make statements like "I actually consumed bundle X in period 1; I would be happier if I *had* consumed X^* and were now to follow it with bundle Y^* than I will be, with my true history, if I now consume bundle Y."

(f) A household-technology version of (e) is obtained if period 1's utility is $u_1(f(L_1), g(G_1), u_2[\bar{f}(L_2, L_1), g(G_2)])$ and period 2's is $u_2[\bar{f}(L_2, L_1), g(G_2)]$, with $u_1, \bar{f}, g, u_2$ increasing in all arguments. As L_1 increases, leisure productivity in period 2 increases (learning by doing). Now C accepts the subsidy policy, provided it increases the chosen value of L_1.

II. Leisure vs. Goods Policies in Many-Person Economies

The study of policies which change leisure-goods choices only becomes interesting in many-person economies. Those who claim that a major shift towards leisure means disaster appear to feel, without benefit of analysis, that people who do not shift would be injured. The preceding one-person explorations suggest that the simplest theoretical treatment of this complex issue is a household-technology treatment. The simplest case worth studying has two persons with different utility functions (easily extendable to two homogeneous groups of persons). Before the policy, person i, $i = (1,2)$, has the function $u_i(f(L_i, \lambda_i), g(G_i))$, as in case (b) above. The production function h converts total time (of both persons) not retained as leisure into a total quantity of goods available for both. An *allocation mechanism* determines who gets what; it assigns a value of (L_1, L_2, G_1, G_2) to each pair (λ_1, λ_2). One example is a price-taking (competitive) mechanism, well-defined only when there are increasing costs ($h' > 0$, $h'' < 0$), which determines $(\bar{L}_1, \bar{L}_2, \bar{G}_1, \bar{G}_2)$ such that for some $p > 0$ (a goods price, the price of labor being one), $ph(48 - L) - (48 - L)$ is maximized at $L = \bar{L}_1 + \bar{L}_2$; $\bar{G}_i = (24 - \bar{L}_i)/p$, $i = 1,2$; and $\bar{L}_i$, $i = 1,2$, maximizes, subject to $L_i \leq 24$, $u_i(f(L_i, \lambda_i), g[(24 - L_i)/p])$. Another mechanism, which may be usable if costs are nonincreasing, is a monopolistic mechanism, determining $(\hat{L}_1, \hat{L}_2, \hat{G}_1, \hat{G}_2)$ such that some $\bar{p} > 0$ (an equilibrium monopoly price) is a maximizer of $ph(48 - \hat{L}_1(p) - \hat{L}_2(p)) - (48 - \hat{L}_1(p) - \hat{L}_2(p))$, where $\hat{L}_i(p)$ maximizes $u_i(f(L_i, \lambda_i), g[(24 - L_i)/p])$; $\hat{L}_i = \hat{L}_i(\bar{p})$, $i = 1,2$; and $\hat{G}_i = (24 - \hat{L}_i)/\bar{p}$. Given standard concavity and differentiability, one readily establishes a

PROPOSITION: *If there are increasing labor costs (in goods production), then, under the price-taking mechanism, a small increase in 1's leisure productivity—in λ_1—increases both persons' utility. If there are nonincreasing costs (so that the price-taking mechanism cannot be used), then existence of a monopoly equilibrium price (permitting use of the monopolistic mechanism) is consistent with a rise in both persons' utility or with a fall in both persons' utility, following a rise in λ_1.*

In this simplest of models, then, there is no a priori theoretical case for the view that if some persons shift towards leisure, others must be hurt: in a smooth increasing-cost competitive world, all benefit; in a nonincreasing-cost world, without specifying much more, the possibility that all benefit remains open.

If one takes this class of models seriously, then they provide clues as to the complicated and novel empirical paths one ought to travel in studying the welfare effects of a shift toward leisure. The leisure and goods commodities produced by the functions f and g need to be identified and measured, using surrogate observables; the structure of preferences over combinations of the two commodities needs to be studied; evidence that certain policies increase leisure productivity for large groups of people needs to be found. As for the economy's nonhousehold sector, the question is whether the economy is presently in the "increasing" or the "nonincreasing cost" range (or an appropriately generalized form of this question once one turns to a multisector economy, replacing the one-dimensional goods variable with a many-dimensional one).

III. Conclusion

The household-technology approach to taste changes has strong appeal for study of the leisure-shift issue; that is likely to be true also for normative analysis of the other two difficult policy issues mentioned at the start. One would not expect analogues of the above relatively clean and simple proposition if one insists on endogenous taste change models and says that a policy improves welfare if everyone's utility some periods hence, or in a steady state, is higher with than without the policy; or if one gives no structure to the utility function, and studies policies which change taste exogenously, knowing only the direction in which they alter marginal rates of substitution of goods for leisure. One would, moreover, have to accept the intrapersonal comparisons we discussed above, or the dismissal of our person C's views as unimportant.

For normative as opposed to predictive questions, household-technology models can be more ambitious than those so far proposed, *provided* one is willing to seek evidence outside of economics. The central question is whether given a proposed policy which initially appears taste altering, one can find a plausible model of a typical individual whose choice set and choice function are unaffected by the policy and whose utilities with and without the policy can be compared. Is there, in fact, an underlying unchanging person, whose basic preferences are his for life, though he may learn unceasingly how to fulfill them better?

To search for models of such a person, one might indeed, as the Becker approach has, view the individual as a factory with joint production of a constant collection of commodities. However there are a variety of ways to model technologies, and none of them need be excluded. They include activity analysis; models wherein part of present production is devoted to "investment" which increases future productivity; models with learning by doing; and models in which production choices are inefficient (to capture an individual's limited rationality).

Price, income, and demand data will never supply the parameters of such models. Psychological inquiry may. The motivational psychology studies cited by Scitovsky, longitudinal studies of individuals who learn over time how to make themselves contented, "quality of life" surveys which try to find the weights people give certain dimensions of their lives—these long and arduous efforts may ultimately provide structure and parameters for usable models. More detailed speculation would be idle; but one has only to glance at some of the existing work of this sort to be convinced that if ever an economist—motivated by the policy and modeling issues we have considered—is asked to help guide the research, it will take a distinctly different turn.[1]

The study of policies which appear to be taste altering, like policies which shift some people toward different uses of their time, can be pushed aside, as illegitimate or unimportant or hopeless. If it is not, then the long-deferred full partnership of economists and noneconomists has to begin.

[1] An example is the recent massive survey study by Angus Campbell et al. To an outsider this appears to be what economists once called "measurement without theory." Had there been some underlying model of an individual, for which the survey responses were to provide parameters, the study would have been organized much differently.

REFERENCES

Angus Campbell et al., *The Quality of American Life*, New York 1976.

P. Hammond, "Changing Tastes and Coherent Dynamic Choice," *Rev. Econ. Stud.*, Feb. 1976, *43*, 159–74.

Staffan Linder, *The Harried Leisure Class*, New York 1970.

E. J. Mishan, *The Economic Growth Debate: An Assessment*, London 1977.

Tibor Scitovsky, *The Joyless Economy*, New York 1976.

C. C. von Weizsäcker, "Notes on Endogenous Change of Tastes," *J. Econ. Theory,* Dec. 1971, *3,* 345–72.

B. Weisbrod, "Comparing Utility Functions in Efficiency Terms, or What Kind of Utility Functions Do We Want?," *Amer. Econ. Rev.,* Dec. 1977, *67,* 991–95.

M. E. Yaari, "Endogenous Changes in Tastes: A Philosophical Discussion," res. memo. no. 23, Center Res. Math. Econ. Game Theory, Hebrew Univ., Oct. 1976.

The Current Experience with Floating Exchange Rates: An Appraisal of the Monetary Approach

By JOHN F. O. BILSON*

In their seminal contributions to the monetary approach to the balance of payments, both Robert Mundell and Harry Johnson suggested that the fixed rate analysis could easily be extended to a world with flexible exchange rates. Their viewpoint is illustrated in the following simple model:

$$M = PK(i, Y) \tag{1}$$

$$M^* = P^*K(i^*, Y^*) \tag{2}$$

$$P = SP^* \tag{3}$$

where M = the money supply; P = the price level; i = the (exogenous) nominal rate of interest; Y = the (exogenous) level of real income; and an asterisk denotes the foreign country. The first two equations relate the demand for real money balances $K(i, Y)$, negatively to the nominal rate of interest and positively to the level of real income. The third equation—the purchasing power parity condition—relates the exchange rate S to the ratio of the nominal price levels. If the foreign country is large relative to the home country, the second equation determines the foreign price level. Under fixed exchange rates, the third equation determines the domestic price level, thus leaving the first to determine the domestic money stock. Under flexible exchange rates, the money supply is exogenous and the first equation therefore determines the domestic price level, while the third is left to determine the exchange rate. Consequently, the particular exchange rate regime decides the set of dependent variables, but does not alter the underlying structure of the model.

The solution for the exchange rate from this quantity theory model is

$$S = \frac{M}{M^*}\frac{K(i^*, Y^*)}{K(i, Y)} \tag{4}$$

The exchange rate, as Jacob Frenkel (1976) has stressed, is the relative price of two moneys. Equation (4) states that the relative price of two moneys is determined by the relative supplies of, and demands for, the two moneys. The equation also yields three important propositions: (a) that an increase in the money supply will result in a proportional depreciation of the exchange rate; (b) that an increase in real income will appreciate the exchange rate; (c) that an increase in the nominal rate of interest will depreciate the exchange rate. Much of the interest in the monetary approach arises from the fact that these propositions are different from the propositions derived from traditional balance-of-payments analysis.

Although there is little doubt of the validity of the monetary model as a long-run theory of exchange-rate determination, certain obvious problems arise when the simple quantity theory model is confronted with the evidence from the current experience with floating exchange rates. First, a number of studies have demonstrated that the purchasing power parity condition does not hold in the short run. This failure is particularly evident when aggregate price indices of the type conventionally used to deflate nominal money balances are used in the calculation. Second, although real income may be argued to be independent of

*International Monetary Fund (*IMF*) and Northwestern University. I am grateful to Rudiger Dornbusch, Jacob Frenkel, and to colleagues at the *IMF* for comments on an earlier draft. The views expressed are solely my responsibility.

the current value of the exchange rate, and although the money supply may be assumed to be an exogenous policy instrument, the assumption that nominal interest rates are exogenous is clearly unjustified. Because international financial markets are highly integrated and efficient, nominal interest rate differentials primarily reflect exchange rate expectations, which are likely to be influenced by the current change in the spot rate. Unless this link between exchange rates, interest rates, and the exogenous variables is specified, the usefulness of the model for policy and forecasting purposes is limited. The third characteristic of the recent experience is the poor performance of the forward rate as a forecast of the future spot rate. This performance does not appear to be due to market inefficiency, since alternative forecasts have not proven superior (see the author and Richard Levich). It should more correctly be attributed to the unpredictable nature of exchange rate movements.

In response to these conditions, a number of new theories have been constructed which attempt to account for the weaknesses in the quantity theory model. The characteristic feature of these models is the emphasis on the role of asset markets in the determination of the exchange rates. Within the general approach, a number of distinct models have been constructed. In what follows, I shall briefly describe three of these models and then attempt to empirically distinquish between them.

The equilibrium rational expectations (*ERE*) version of the monetary approach (see the author, 1978a; Frenkel, 1976; Robert Hodrick; Michael Mussa) argues that the failure of the purchasing power parity condition is due to the fact that published price indices are poor indicators of transactions prices and hence are inappropriate deflators for nominal money balances. Stephen Magee has demonstrated that contractual and recording conventions lead to spurious deviations from purchasing power parity between official price indices which may not be present in transactions prices. On these grounds, the *ERE* approach assumes that the "true" price indices are unobservable variables whose ratio is defined to be equal to the exchange rate. Eliminating the unobservable variables in equations (1)-(3) leads directly to equation (4).

The *ERE* theory of the interest rate differential proceeds in three stages: first, the forward exchange market is assumed to be dominated by rational, risk-neutral speculators who purchase all forward contracts at a forward rate equal to the expected future spot rate. Second, covered interest arbitrage is assumed to set the interest rate differential equal to the forward exchange premium. Finally, expectations are assumed to be rational: the expected rate of depreciation is set equal to the rate predicted by the model itself. The solutions to these models relate the interest rate differential to the actual and expected future values of the exogenous variables. It is shown that expected future monetary growth is "discounted" into the current spot exchange rate in an analogous manner to the way in which revisions in expected future earnings are discounted into equity prices. The cause of volatility in the foreign exchange market is the instability in the underlying process generating the exogenous variables.

The currency substitution models of Guillermo Calvo and Carlos Rodriguez; Lance Girton and Don Roper; David King, Bluford Putnam and Sykes Wilford; Arthur Laffer; and Marc Miles constitute a second approach to the stylized facts of the floating rate period. The currency substitution (*CS*) models emphasize that real money balances as an aggregate may be held in a portfolio of currencies, so that the theory of the exchange rate is a problem in portfolio selection. Since *CS* models are typically based upon an integrated economy, be it an individual country or the world, the purchasing power parity problem is circumvented by expressing the real value of all currencies in terms of their purchasing power over a common bundle of goods. For the same reason, these models deemphasize the role of relative real incomes as a determinant of the exchange rate. These assumptions clear the way for the relative supplies

of the two currencies, and the interest rate differentials, which measure exchange rate expectations, to play the central role in the analysis. In the *CS* models, the instability of the exchange rate is more likely to be attributed to a high elasticity of substitution between currencies. In the limit case of almost perfect substitutability, the holding cost on all coexisting currencies must be the same, the equality being brought about by instantaneous shifts in the spot rate. Empirically, this would imply that exchange rates follow a random walk with zero drift. More generally, if the elasticity of substitution is high, slight changes in money growth rates will lead to substantial changes in the spot exchange rate, and a persistent inflationary policy may drive a currency into hyperinflation. The freedom to conduct an independent monetary policy, which is often cited as the major advantage of flexible exchange rates, is an illusion in a world of currency substitution.

The final interpretation of the current experience is contained in Rudiger Dornbusch's disequilibrium model of exchange rate dynamics. Unlike the *ERE* and *CS* models, the Dornbusch model assumes that the slowly adjusting commodity price indices are the appropriate deflators for nominal money balances. With prices fixed in the short run, a monetary expansion must be accompanied by a fall in the nominal rate of interest, but this implies from the interest rate parity condition that the spot rate must depreciate by more than the forward rate. Rational expectations then requires that the spot rate *appreciate* during the period of price adjustment, so that the rate must have initially depreciated by more than in proportion to the increase in the money supply. Hence the cause of exchange rate instability in the Dornbusch model is the differential speed of adjustment between commodity and asset markets.

Each of the preceding models provide internally consistent explanations of the current experience: each accounts for the slow adjustment of commodity prices; the sensitivity of exchange rates and interest rates to money market conditions; and the efficiency of international capital markets. However, the policy implications of the models are quite different. The *ERE* approach stresses constant money growth rate rules, the *CS* models suggest a need for international monetary coordination, and the Dornbusch model may be used to justify a more active policy of exchange rate management.

There are three grounds upon which an empirical distinction can be made between the three models: first, the *ERE* and *CS* models suggest that the exchange rate is the appropriate deflator for the relative money supplies, while the Dornbusch model argues that the ratio of domestic prices is more appropriate. Secondly, the *CS* models suggest that relative real incomes will not have a significant impact on the exchange rate, and that the interest elasticity of the demand for money is higher during the floating rate period because of the wider possibilities for currency substitution. Finally, the *ERE* models predict a positive relationship between the money supply and the interest rate, while the Dornbusch model suggests a negative relationship. The preceding overstates the differences between the three approaches, but this is an unavoidable necessity in any empirical test.

These predictions were examined in an empirical model, which is estimated using data on the deutschemark/pound exchange rate over the period from June 1970–August 1977. This rate was chosen because the variables stressed by the monetary approach—relative real incomes, relative money supplies, and interest rate differentials—have varied sufficiently over the sample period for their influence upon the exchange rate to be gauged. A detailed discussion of the data sources is given in the author (1978b). Briefly, the "price index" is the consumer price index, the "money supply" is defined to be the M_2 definition for Germany and the M_3 definition for the United Kingdom, real income is proxied by the seasonally adjusted industrial production index, and the interest rate differential is proxied by the one-month forward premium. The model,

estimated by two-stage least squares, is presented in the following three equations:[1]

$$s_t = -.425\,(p_t - s_t) + 11.376\,x_t$$
$$\quad\;(3.0) \qquad\qquad (3.8)$$
$$+\; 1.264\,m_t - 1.385\,y_t$$
$$(14.9) \qquad (7.1)$$
$$+\; 2.027\,DV_t + u_{1t}$$
$$(5.3)$$

R^2 (adj.) = .995; S.E. = .0608; D.W. = 1.651; Estimate of first-order autocorrelation coefficient = .506

$$Dp_t = .020\,(s_{t-1} - p_{t-1})$$
$$(2.8)$$
$$+\; .291\,Dp_{t-1} + u_{2t}$$
$$(2.8)$$

R^2 (adj.) = .2109; S.E. = .0067; D.W. = 2.046

$$Dx_t = .922\,(\underline{x}_t - Ds_t)$$
$$(11.4)$$
$$+\; .893\,(Ds_t - x_{t-1}) + u_{3t}$$
$$(11.8)$$

R^2 (adj.) = .6187; S.E. = .0013; D.W. = 1.373

In this presentation, the following definitions are employed: $s = ln(S)$; $p = ln(P/P^*)$; $x = i - i^*$; $m = ln(M/M^*)$; $y = ln(Y/Y^*)$ and D denotes the difference operator. The DV_t variable in the first equation is a dummy variable whose purpose is to capture the effect of the oil revenues on the demand for sterling. It is equal to .15 in January 1974 and declines exponentially thereafter. The parenthesized numbers are t-statistics.

The first equation is a *log*-linear version of equation (4). The first term determines the appropriate deflator for the relative money supplies: if the coefficient is minus unity, the relative consumer prices are the appropriate deflator while the exchange rate is appropriate if the coefficient is zero. The estimated ceofficient of −.425 suggests that an intermediate position is correct. It is certainly not possible to ignore conventional price indices in the money supply deflator. The coefficient on the interest rate differential demonstrates the existence of a strong and significant relationship between this variable and the exchange rate. However, since the implied interest elasticity—approximately .1 when the nominal rate of interest is 10 percent per annum—is low relative to other estimates, there does not appear to be any support for the view that the elasticity of substitution has increased during the floating rate period. Nor is there any support for the prediction that relative real incomes do not influence the exchange rate: the implied income elasticity of 1.385 is both strong and statistically significant. Finally, the elasticity relating the exchange rate to the money supply is significantly greater than unity, even without allowance for distributed lags. Hence, despite the allowance made for other channels of influence, the exchange rate remains sensitive to monetary disturbances.

The second equation relates the relative rate of inflation to the deviation from purchasing power parity. The estimates support the view that relative consumer prices gradually adjust to the exchange rate. It takes twelve months to complete one-quarter of the necessary adjustment of prices to an exchange rate change, twenty-seven months to complete one-half of the adjustment and fifty-four months to complete 75 percent. It is extremely difficult to believe that these long lags are simply the result of contractual arrangements.

The final equation makes use of the two-part expectations mechanism developed by Frenkel (1975) to explain the change in the interest rate differential. Through the first term, an increase in the actual rate of depreciation above the long-term expected rate $\underline{x}_t$ (which is proxied by the twelve-month forward premium) will cause a decrease in the short-term expected rate, because speculators will anticipate a return to

[1] A constant term was allowed in each of the three equations, although the estimated coefficient is not reported in the text. The estimates of the first equation are based upon a weighted Cochran-Orcutt transformation of the original equation, because the residuals from that equation exhibited strong patterns of serial correlation and hetroscedasticity. The weighting factor is a five-period moving average of the squared errors of the whitened residuals. Further details on the estimation procedures are available from the author upon request.

the long-term rate. However, through the second term, an increase in the actual rate above the previous periods short-term rate will cause an increase in the short-term expected rate. In the deutschemark/pound case, the regressive element dominates the second adaptive element. This implies that a depreciation of the exchange rate will be associated with a decline in interest rates, as suggested by the Dornbusch model. It is, however, possible that a positive relationship exists between the long-term rate and the money supply, as suggested in the *ERE* model. The results should therefore be interpreted cautiously.

However, it is true that the estimates tend to confirm the predictions of the Dornbusch model. As it suggests, an increase in the money supply will cause an overshooting in the exchange rate and a decrease in interest rates in the short run, and the medium-term dynamics of the model do appear to be dominated by the slow adjustment of commodity prices. Since this suggests that the exchange rate or, more correctly, monetary policy does have real effects, the logical extension of the model is to take account of the induced movement in relative real incomes.

The results also offer strong support for the monetary approach in general. The illustrative model presented above is extremely simple, yet it accounts for over 99 percent of the variance in the volatile deutschemark/pound rate. All of the important elasticities are of the correct sign and are statistically significant, and the model appears to take account of all of the problems that arose in the application of the simple quantity theory model to the recent experience. While there is still much work to be done, it appears that the marriage of theory and evidence, which appeared precarious in the early years of the float, has strengthened over time as theory has developed and evidence accumulated.

REFERENCES

J. F. O. Bilson, (1978a) "Rational Expectations and the Exchange Rate," in Jacob Frenkel and Harry Johnson, eds., *The Economics of Exchange Rates: Selected Studies*, Reading 1978.

———, (1978b) "The Monetary Approach to the Exchange Rate: Some Empirical Evidence," *Int. Monet. Fund Staff Pap.*, Mar. 1978.

——— **and R. M. Levich,** "A Test of the Forecasting Efficiency of the Forward Exchange Rate," unpublished manuscript, New York Univ. 1977.

G. A. Calvo and C. A. Rodriguez, "A Model of Exchange Rate Determination with Currency Substitution and Rational Expectations," *J. Polit. Econ.*, June 1977, *85*, 617-26.

R. Dornbusch, "Expectations and Exchange Rate Dynamics," *J. Polit. Econ.*, Dec. 1976, *84*, 1161-76.

Jacob A. Frenkel, "Inflation and the Formation of Expectations," *J. Monet. Econ.*, Oct. 1975, *1*, 403-21.

———, "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scand. J. Econ.*, No. 2, 1976, *78*, 200-24.

——— **and Harry G. Johnson,** *The Monetary Approach to the Balance of Payments*, London 1975; Toronto 1976.

——— **and** ———, *The Economics of Exchange Rates: Selected Studies*, Reading 1978.

L. Girton and D. Roper, "Theory and Implications of Currency Substitution," disc. paper no. 56, Fed. Res. Board, Washington 1976.

R. J. Hodrick, "An Empirical Analysis of the Monetary Approach to the Exchange Rate," in Jacob Frenkel and Harry Johnson, eds., *The Economics of Exchange Rates: Selected Studies*, Reading 1978.

H. G. Johnson, "The Monetary Approach to the Balance of Payments," in Jacob Frenkel and Harry Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1975; Toronto 1976.

D. T. King, B. H. Putnam, and D. S. Wilford, "A Currency Portfolio Approach to Exchange Rate Determination," unpublished manuscript, Fed. Res. Bank New York, July 1977.

A. Laffer, "Optimal Exchange Rates," unpublished manuscript, Univ. Chicago 1976.

S. P. Magee, "Contracting and Spurious Deviations from Purchasing Power Parity," in Jacob Frenkel and Harry Johnson, eds., *The Economics of Exchange Rates: Selected Studies*, Reading 1978.

M. Miles, "Currency Substitution, Flexible Exchange Rates, and Monetary Independence," unpublished manuscript, Rutgers Univ. 1976.

Robert A. Mundell, *International Economics*, New York 1968.

M. Mussa, "The Exchange Rate, the Balance of Payments, and Monetary and Fiscal Policy under a Regime of Controlled Floating," *Scand. J. Econ.*, No. 2, 1976, *78*, 229–48.

New Views of Exchange Rates and Old Views of Policy

By PETER B. KENEN*

In what sense can it be said that a floating exchange rate confers national autonomy? Can it insulate an economy against disturbances coming from abroad? Can it enhance the effectiveness of domestic monetary and fiscal policies? How does the degree of capital mobility influence our answers to these questions? Experience with limited exchange rate flexibility since 1973 would appear to say that floating rates cannot confer complete autonomy. Furthermore, developments in the theory of exchange rate determination, inspired in part by that experience, have given us new ways of approaching these newly controversial issues.

I propose to deal with these questions in two ways. First, I will present results extracted from a formal algebraic model on which I have been working in collaboration with Polly Allen and which forms the basis for our forthcoming book. Second, I will summarize results obtained from an expanded version of an econometric model first presented in my 1974 article and soon to appear in my forthcoming book.

I

The familiar questions with which I began have to be answered anew because the models used in current work on international theory include phenomena that did not figure in older models. In particular, the current models employ a portfolio or asset-market approach and allows for wealth effects of exchange rate changes in the determination of the balance of payments and the exchange rate. This approach subsumes the more narrowly conceived monetary approach to the balance of payments.

The model from which I draw my findings is summarized in Table 1 and has features found in many other models (see, especially William Branson; Lance Girton and Dale Henderson; Rudiger Dornbusch, 1977; Russell Boyer). The model describes a small open economy facing fixed prices for the goods and bonds it *buys* abroad, but not for those it *sells* abroad. Capital movements take place exclusively by way of transactions in foreign bonds denominated in foreign currency. The economy is always in equilibrium; all markets are perfectly competitive and clear continuously; and expectations are perfectly stationary. (For models in which expectations figure prominently in the determination of a floating exchange rate, see Dornbusch, 1976; Pentti Kouri.)

The economy produces only one good Q_1, and output is an increasing function of the home currency price of the good p_1. Thus the country's gross domestic product Y is price *times* quantity, as at (1) in Table 1.

Household wealth W^h is held as money L^h, domestic bonds denominated in home currency B^h, and foreign bonds denominated in foreign currency F^h. (Both bonds are bills, so that capital gains and losses arise only because the home currency value of a foreign bond depends on the exchange rate π.) Wealth is written as the sum of the histories of saving S, and of capital gains and losses on foreign bonds, this being the principal dynamic relationship in the model.

Saving is made to depend on foreign and home interest rates, $\bar{r}_0$ and r_1, on disposable income, Y^d, and on wealth. Disposable income is gross domestic product *plus* interest income earned on bond holdings *less* lump sum taxes T^h, paid by households. Consumption C is disposable income *less* saving, and the demands for the two goods, C_0 and C_1, depend on consumption and on the home currency

*Princeton University.

TABLE 1—THE ALGEBRAIC MODEL

(1) The Supply Side

$$Y = p_1 Q_1, Q_1 = f(p_1), f_1 \geq 0^a$$

(2) Wealth, Saving, and Demands for Goods

$$W^h = L^h + B^h + \pi F^h = \int S\,dt + \int F\left(\frac{\delta\pi}{\delta t}\right) dt$$

$$S = S(\bar{r}_0, r_1, Y^d, W^h), S_0 > 0, S_1 > 0, 0 < S_Y < 1, S_W < 0$$

$$Y^d = Y + \bar{r}_0(\pi F^h) + r_1(B^h) - T^h = C + S$$

$$C = \pi\bar{p}_0^f C_0 + p_1 C_1, C_1 = C_1(\pi\bar{p}_0^f, p_1, C), C_{10} > 0, C_{11} < 0, C_{1C} > 0^b$$

$$C_1^f = C_1^f\left(\bar{p}_0^f, \frac{p_1}{\pi}, \bar{C}^f\right), C_{10}^f > 0, C_{11}^f < 0, C_{1C}^f > 0$$

(3) Demands for Assets

$$L^h = L(\bar{r}_0, r_1, Y, W^h), L_0 < 0, L_1 < 0, L_Y > 0, L_W > 0$$

$$B^h = B(\bar{r}_0, r_1, Y, W^h), B_0 < 0, B_1 > 0, B_Y < 0, B_W > 0$$

$$\pi F^h = F(\bar{r}_0, r_1, Y, W^h), F_0 > 0, F_1 < 0, F_Y < 0, F_W > 0^c$$

(4) The Central Bank, Money, and Exchange Rate Policy

$$L = \bar{B}^c + \pi R, R = 0 \text{ or } \pi = \bar{\pi}$$

(5) The Government, Fiscal Policy, and Supply of Domestic Bonds

$$D = \bar{G}_1 + r_1(B - \bar{B}^c) + T^f - T^h$$

$$B = \int D\,dt$$

$$T^f = \bar{r}_0(\pi F^h)$$

(6) The Market-Clearing Equations

$$p_1 C_1 + p_1 C_1^f + \bar{G}_1 - p_1 Q_1 = p_1 C_1^f - \pi\bar{p}_0^f C_0 + \bar{D} - S = 0$$

$$B^h + \bar{B}^c - B = 0$$

$$L^h - L = 0$$

[a] If output depends on labor input and diminishing returns prevail, the simplest Keynesian case can be obtained by assuming that labor supply is perfectly elastic at a fixed money wage (in which case $f_1 > 0$) and the simplest classical case can be obtained by assuming that labor supply is perfectly inelastic and the money wage rate perfectly flexible (in which case $f_1 = 0$).

[b] The demand functions for goods are assumed to be homogeneous of degree zero in prices and nominal consumption, and to have unitary elasticities with respect to nominal consumption.

[c] As demands for assets are constrained by actual wealth, it can be shown that $L_W + B_W + F_W = 1$, $L_Y + B_Y + F_Y = 0$, and $L_i + B_i + F_i = 0$, for $i = 0, 1$.

prices of the two goods, $\pi\bar{p}_0^f$ and p_1. The foreign demand for the domestic good C_1^f, is defined analogously, using the relevant foreign currency arguments.

The households' demands for money, domestic bonds, and foreign bonds are constrained by wealth and are written in nominal terms as functions of interest rates, income, and wealth, as at (3) in Table 1.

Domestic money is issued by the central bank. Thus, the stock of money L is defined at (4) as the sum of the central bank's holdings of domestic bonds $\bar{B}^c$, and of foreign currency reserves R. The central bank adjusts its holdings of domestic bonds by open-market operations. It adjusts its holdings of reserves to execute exchange rate policy. Under a floating rate, it abstains completely from intervention in the foreign exchange market and has no need for reserves. It is thus convenient (but not necessary) to suppose that the stock of reserves is zero. Under a pegged rate, the central bank intervenes to guarantee that the rate remains at the desired level $\bar{\pi}$, and its holdings of reserves vary accordingly.

The government buys domestic goods; its demand $\bar{G}_1$ is policy determined in nominal terms. The government's budget constraint is given at (5) in Table 1, where D is its nominal budget deficit, B is the stock of government debt (the supply of domestic bonds) and is determined by the history of budget deficits, and T^f are transfers to foreigners.

If the budget deficit were endogenous as in most macroeconomic models, there could be no clear-cut distinction between

goods-market disturbances and asset-market disturbances; any disturbance or policy change impinging on any term in the budget equation would affect the deficit and the supply of bonds. For this and other reasons, it is useful to suppose that the budget deficit is policy determined. The government selects a deficit of predetermined size and duration, achieving its aim by adjusting lump sum taxes T^h continuously. By implication, long-run changes in the stock of debt B are likewise policy determined, being the steady-state counterparts of temporary deficits. Finally, I assume that the government adjusts transfers to foreigners continuously to offset exactly the interest income that households earn from foreigners. This assumption causes the current account balance to equal the trade balance and simplifies the working definition of disposable income.

Under my version of the small country assumption, the supply of foreign goods is perfectly elastic at the foreign currency price $\bar{p}_0^f$, and the supply of foreign bonds is perfectly elastic at the foreign interest rate $\bar{r}_0$. We have thus to write down and solve simultaneously only three market-clearing equations—for the domestic good, the domestic bond, and domestic money, as at (6) in Table 1.

II

Using these three equations, the model can be solved for the market-clearing values of p_1, r_1, and π (or R), given the exogenous and policy variables and two state variables—the stock of domestic debt and the integral of household saving. (Wealth itself is not a state variable because it is affected instantaneously by a change in the exchange rate.) And because the model is stable dynamically, it can also be solved for the steady-state values of those same variables—those that obtain when saving goes to zero.

Before returning to the questions raised at the start of this paper, let me draw attention to three points:

1) The exchange rate is the price of one national money in terms of another. Accordingly, monetarists are fond of saying that the rate is determined in and by the money market, given the effects of disturbances and policies on all other markets. This view is misleading. In my model, as in others, the exchange rate is the price that *clears* the money market, but it is determined jointly with income and the interest rate by the responses of all markets together. A change in the exchange rate affects the home currency prices of foreign goods and of foreign bonds and has therefore to influence the goods and bond markets. In my model, moreover, it affects household wealth and by this route affects saving and absorption.

2) Price elasticities appear pervasively in the solutions. This fact likewise contradicts the simplistic view that the exchange rate is determined in and by the money market. The real or barter side of the economy is *not* irrelevant to the determination of a floating exchange rate. It is relevant, in fact, even to the effects of disturbances arising *in* the money market. (The sizes of the changes in π and Y resulting from an open-market operation depend on goods-market elasticities.)

3) The signs of many outcomes depend on an assumption that the absorption-increasing effects of a change in wealth are not "crowded out" by the absorption-decreasing effects of the concomitant change in the interest rate. To be precise, I assume that saving (absorption) is relatively sensitive to changes in wealth, while the demand for money is relatively sensitive to changes in the interest rate, so that $S_W \cdot L_1 > L_W \cdot S_1$. This is vital, for example, to my finding that an increase in the stock of debt, the long-run counterpart of a budget deficit, raises the steady-state level of income. (It plays the same role at that point as an analogous assumption made by Alan Blinder and Robert Solow.) It is likewise important to my finding that an increase in the foreign interest rate has a positive effect on income. In that instance, however, I add the supposition that the effects of changes

in $\bar{r}_0$ and r_1 on the demand for money are proportional to their effects on saving. Finally, it plays a strategic role in deciding the influence of capital mobility on the effectiveness of fiscal and monetary policies.

III

What does this model have to say about the ability of a floating exchange rate to insulate the national economy from an external disturbance? Consider first the signs of the effects of an increase in the foreign price. There is an immediate and permanent appreciation of the domestic currency and a temporary increase in income. Thus, insulation is not instantaneous but occurs with the passage of time, as saving and capital flows lead to changes in stocks of wealth and holdings of foreign bonds that cause income to return to its initial level. To put the point generally, in the very short run, a floating exchange rate has the task of clearing the money market. Gradually that task is taken over by changes in wealth, income, and the interest rate. Then the floating exchange rate takes on a different task; it clears the goods market, given the level of spending implied by the steady-state level of income, by acting on the trade balance.

It can thus be said that insulation is achieved, but only in the long run. This is true for a large class of goods-market disturbances, including some that have *domestic* origins. It is achieved against an increase in foreign expenditure, $\bar{C}^f$, but also against a shift of domestic or foreign demand between domestic and foreign goods.

Insulation is not achieved, however, against a foreign asset-market disturbance, not even in the long run. An increase in the foreign interest rate causes the domestic currency to depreciate permanently and increases income. (This result is counterintuitive, as an increase in $\bar{r}_0$ raises saving and is to this extent *de*flationary. However it also raises the domestic demand for the foreign bond, and therefore the demand for foreign currency, so that the domestic currency depreciates, raising the home and foreign demands for the domestic good. There is thus a trade balance effect that swamps the direct absorption effect.)

Returning to the process of insulation against goods-market disturbances, let us see why it is not instantaneous. Consider the effects of a goods-market disturbance that raises domestic output and income, and therefore the demand for money. As the supply of money cannot respond endogenously under a floating rate, the home currency must appreciate to reduce the demand for money, and there are two ways in which it does so. First, the trade balance worsens, reducing income. Second, households suffer capital losses on their holdings of foreign bonds, reducing wealth. If households held no such bonds, the trade balance effect would have to do the whole job of clearing the money market, and the exchange rate would have to appreciate sufficiently to offset completely the income raising effect of the initial disturbance. When they do hold foreign currency bonds, however, the capital loss effect is called into play and reinforces the decline in the demand for money. The trade balance does not have to offset the entire income raising effect of the disturbance.

IV

What does this model have to say about the powers of fiscal and monetary policies under a floating exchange rate and about the influence of capital mobility? Answers are supplied by Table 2. Let us begin with fiscal policy, represented in the short run by an increase in the budget deficit $\bar{D}$, and in the long run by an increase in the stock of debt $\bar{B}$. In the short run it raises output and income, even in the case of "perfect" capital mobility—when foreign and domestic bonds are perfect substitutes. It can be shown, however, that the short-run effect of a budget deficit is *smaller* with a floating exchange rate than with a pegged rate, regardless of the degree of capital mobility.

TABLE 2—THE INFLUENCE OF CAPITAL MOBILITY ON THE EFFECTIVENESS OF MONETARY AND FISCAL POLICIES UNDER FLOATING AND PEGGED EXCHANGE RATES

	Floating Rate		Pegged Rate	
Outcome	Impact Effect	Steady State	Impact Effect	Steady State
Sign of Income Change				
With Imperfect Substitutability Between Bonds				
Budget deficit	+	+[a]	+	0
Open-market purchase	+	+	+	0
With Perfect Substitutability Between Bonds				
Budget deficit	+	0	+	0
Open-market purchase	+	+	0	0
Effect of Increase in Substitutability on Size of Income Change				
Budget deficit	Decreases	Decreases[a]	Increases	None
Open-market purchase	Increases[b]	Increases	Decreases	None
Effect of Increase in Domain of Foreign Bond on Size of Income Change				
Budget deficit	Increases	None	None	None
Open-market purchase	Ambiguous[c]	None	None	None

[a]On the assumption that crowding out does not dominate.
[b]The assumption that crowding out does not dominate is sufficient, but not necessary for this result.
[c]Will decrease the size of the increase in income in the case in which an open-market purchase causes the exchange rate to depreciate.

These results are new. Robert Mundell argued that with no capital mobility fiscal policy is more effective under a floating rate. In his model, however, capital mobility diminishes its influence under a floating rate while increasing its influence under a pegged rate. In the limiting case of perfect mobility, fiscal policy is utterly ineffective under a floating rate. While the signs of the effects of an increase in capital mobility are the same in my model as in Mundell's, the presence of foreign-currency bonds prevents an increase in capital mobility from depriving fiscal policy of all its influence under a floating rate. The capital loss effect limits the change in the exchange rate, and the size of the increase in income due to a budget deficit goes to a lower but positive limit as we approach perfect mobility.

There is no need to dwell on the long-run effects of fiscal policy summarized in Table 2. They resemble those obtained from many other models, including the one used by Ronald McKinnon and Wallace Oates when they introduced the vital distinction between the short-run and long-run effects of macro policies. Thus a budget deficit has no permanent effect on income under a pegged exchange rate, and the size of its effect under a floating rate varies inversely with the degree of capital mobility (falling to zero in the long run with perfect mobility).

Turning next to the effects of monetary policy, we encounter new results that do not depend directly on the presence in this model of a foreign currency bond. It can be shown, for example, that monetary policy is not always more effective under a floating rate than under a pegged rate. An open-market purchase is more effective in the short run only when it causes the exchange rate to depreciate—which does not always happen in this model. Furthermore, the degree of capital mobility does not

necessarily raise the short-run effectiveness of monetary policy under a floating rate; we can be sure that it will do so only when we are willing to assume that crowding out does not dominate.

Finally, on the last two lines of Table 2, I explore a neglected dimension of capital mobility—the influence of the domestic *domain* of the foreign currency bond. It has, of course, no relevance for the functioning of policies under a pegged rate, as there are then no capital gains or losses on holdings of foreign bonds. Under a floating rate, however, an increase in the size of the domain of the foreign bond has important implications for the sizes of the short-run effects of domestic policies.

V

What does my econometric model say about the degree of insulation provided by a floating rate and about the influence of the exchange-rate regime on the operations of domestic policies? The model has much in common with the abstract model I have been describing. Foreign currency assets do not appear explicitly in any of the equations, and there are thus no capital gains or losses due to exchange rate changes. Nevertheless, exchange rate changes prove to be influential not only on trade and service flows, but also on capital flows, and there are proxies in the model for expectations of exchange rate changes. Furthermore, capital flows are treated as once and for all adjustments of actual to desired stocks, and significant stock adjustment terms appear in most of the capital flow equations. Finally, trade and other current account flows react with long lags to exchange-rate changes, and these lags have large implications for the ability of a floating rate to insulate the economy from the effects of external disturbances.

The balance-of-payments sector of the model contains more than thirty-five behavioral equations; it is similar in size to the model developed by S. Y. Kwack, although different in specification. The domestic sector contains more than forty behavioral equations; it is a conventional Keynesian model of the *U.S.* economy, but it deals in some detail with the determination of the price level and with financial relationships.

Three pairs of simulations are summarized in Table 3. The first pair is addressed to the question of insulation. It describes the effect of a permanent increase in the level of real economic activity in the outside world (a disturbance resembling an increase in $\overline{C}^f$ in my theoretical model). When the exchange rate is pegged, as in the first simulation, there is an increase of real gross national product in the United States, an improvement in the current account balance, and an increase in the overall balance-of-payments surplus. When the exchange rate floats without official intervention, insulation is incomplete, but it is not inconsequential. The dollar begins to appreciate immediately, limiting the increase in the current account surplus and, therefore, the increase in real *GNP*.

In the second pair of simulations, we ask what happens when income taxes are reduced in the United States (a policy change resembling an increase of $\overline{D}$ in my theoretical model, although tax collections are not adjusted continuously to hold the budget deficit at some desired level). Under a pegged exchange rate, there is an increase in real GNP and a modest deterioration in the current account balance. The overall balance of payments improves initially owing to a reduction in capital outflows. Analogously, the dollar appreciates at first under a floating rate, just as it did in my theoretical model, and the changes in real *GNP* are very slightly smaller in the first three quarters of the simulation. Later, moreover, the dollar depreciates as in my model and the depreciation strengthens the current account balance. As a result, the medium-term effect of a tax cut is larger with a floating rate.

In the third pair of simulations, we ask what happens when the supply of high powered money is made to grow faster in the United States (a policy change resembling an increase of $\overline{B}^c$ in my theoretical model). A large capital outflow reduces the balance-of-payments surplus under a pegged rate and causes the dollar to de-

TABLE 3—EFFECTS OF DISTURBANCES AND POLICY CHANGES UNDER PEGGED AND FLOATING EXCHANGE RATES IN SIMULATIONS USING A QUARTERLY ECONOMETRIC MODEL OF THE *U.S.* BALANCE OF PAYMENTS AND DOMESTIC ECONOMY

	Differences Between Simulations and Control Solutions							
Simulation and Variable	1Q	2Q	3Q	4Q	8Q	12Q	16Q	20Q
Increase in Foreign Activity								
With pegged exchange rate								
Real *GNP*	0.46	0.72	0.82	0.88	1.05	1.04	0.97	0.89
Current account surplus	0.33	0.31	0.31	0.31	0.32	0.30	0.31	0.32
Balance-of-payments surplus	0.04	0.15	0.18	0.97[a]	0.19	0.21	0.21	0.23
With floating exchange rate								
Real *GNP*	0.42	0.53	0.50	0.49	0.44	0.37	0.43	0.41
Current account surplus	0.29	0.21	0.17	0.14	0.09	0.10	0.15	0.14
Composite exchange rate	−0.21	−1.06	−1.60	−1.98	−1.78	−1.88	−2.74	−3.57
Reduction in Income Taxes								
With pegged exchange rate								
Real *GNP*	0.24	0.44	0.63	0.81	1.26	1.44	1.51	1.50
Current account surplus	−0.02	−0.04	−0.05	−0.06	−0.12	−0.19	−0.28	−0.37
Balance-of-payments surplus	0.01	0.01	0.01	−0.02	−0.10	−0.17	−0.24	−0.33
With floating exchange rate								
Real *GNP*	0.24	0.43	0.62	0.82	1.46	1.91	2.20	2.35
Current account surplus	−0.03	−0.04	−0.04	−0.05	−0.09	−0.03	−0.05	−0.08
Composite exchange rate	−0.01	−0.06	−0.03	0.08	1.01	1.60	2.17	3.58
Increase in Growth Rate of Supply of High Powered Money								
With pegged exchange rate								
Real *GNP*	0.01	0.09	0.37	0.78	1.85	3.25	3.24	3.82
Current account surplus	0.00	0.03	0.02	−0.02	−0.07	−0.24	−0.36	−0.56
Balance-of-payments surplus[b]	−0.04	−0.52	−0.57	−0.14[a]	−0.65	−0.22	−0.78	−1.69
With floating exchange rate								
Real *GNP*	0.05	0.64	1.38	1.62	3.32	4.25	4.70	5.40
Current account surplus[c]	0.02	0.38	0.49	0.25	0.49	−0.01	0.19	0.14
Composite exchange rate[c]	0.27	3.16	5.24	3.97	4.91	2.49	8.07	11.87

Note: Data used and actual simulations may be obtained from the author. Exchange rates and foreign activity refer to trade weighted indices.

[a] Reflects a reduction in capital outflows resulting from an endogenous abatement of speculation against the dollar.

[b] Outcomes in certain quarters affected by capital flows resulting from changes in the discount rate.

[c] Outcomes in certain quarters affected by capital flows mentioned in fn. b above; under a floating rate, these affect the current account balance by way of the exchange rate.

preciate under a floating rate. There is an increase in real *GNP* in both instances, but it is much larger, as expected under a floating rate, because the depreciation of the dollar improves the current account balance.

There is thus a remarkable degree of consistency between the signs and relative sizes of the outcomes in these simulations and the predictions summarized in earlier tables. It would be nice to know that this is not an accident—that the econometric model behaves as it does because it captures the chief features of the abstract model. I could point to several fortuitous consistencies, as well as citing many analytical consistencies in the spirit if not the particulars of specification. I therefore draw the usual conclusion that much work remains to be done.

REFERENCES

A. S. Blinder and R. M. Solow, "Does Fiscal Policy Matter?," *J. Publ. Econ.*, Nov.

1973, *2*, 319–37.

R. S. Boyer, "Devaluation and Portfolio Balance," *Amer. Econ. Rev.*, Mar. 1977, *67*, 54–63.

W. H. Branson, "Portfolio Equilibrium and Monetary Policy with Foreign and Non-Traded Assets," in Emil-Maria Claassen and P. Salin, eds., *Recent Issues in International Monetary Economics*, Amsterdam 1976.

R. Dornbusch, "Expectations and Exchange Rate Dynamics," *J. Polit. Econ.*, Dec. 1976, *84*, 1161–76.

———, "Capital Mobility and Portfolio Balance," in Robert Z. Aliber, ed., *The Political Economy of Monetary Reform*, Montclair 1977.

L. Girton and D. Henderson, "Central Bank Operations in Foreign and Domestic Assets under Fixed and Flexible Exchange Rates," in Peter Clark et al., eds., *The Effects of Exchange Rate Adjustments*, Washington 1977.

Peter B. Kenen, *A Model of the U.S. Balance of Payments*, forthcoming.

———, "The Balance of Payments and Policy Mix: Simulations Based on a U.S. Model," *J. Finance*, May 1974, *29*, 631–54.

——— **and Polly R. Allen,** *Asset Markets, Exchange Rates, and Economic Integration*, forthcoming.

P. J. K. Kouri, "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run," *Scand. J. Econ.*, No. 2, 1976, *78*, 280–304.

S. Y. Kwack, "Simulations with a Model of the U.S. Balance of Payments: The Impact of the Smithsonian Exchange-Rate Agreement," in Peter Clark et al., eds., *The Effects of Exchange Rate Adjustments*, Washington 1977.

Ronald I. McKinnon and Wallace E. Oates, *The Implications of International Economic Integration for Monetary, Fiscal, and Exchange-Rate Policies*, Princeton 1966.

Robert A. Mundell, *International Economics*, New York 1968.

Monetary vs. Traditional Approaches to Balance-of-Payments Analysis

By NORMAN C. MILLER*

My objective here is to suggest a perspective on the relationship between the state of the money market and international reserve flows. This will be done by focusing on the monetary approach to the balance of payments (*MBP*) and its relationship to more traditional thinking about the balance of payments. The literature seems to contain two versions of the monetary approach: an *equilibrium MBP* and a *disequilibrium MBP*. The equilibrium version is quite general, but the disequilibrium version requires some special (but reasonable) assumptions. Also, the equilibrium *MBP* suggests a small country offset coefficient (the percent of a home monetary expansion that is dissipated via a loss of international reserves) of minus unity, while the disequilibrium *MBP* yields an offset between minus unity and zero. Furthermore, it is argued that recent papers with an *MBP* flavor reach conclusions that differ from those reached in earlier models, not because of a focus on the money market, but because of different implicit assumptions about which markets clear. Finally, and perhaps most importantly, it is suggested that the state of the money market may bear no necessary relationship to the balance of payments if the government pursues an interest rate pegging-type of monetary policy, or if there is a nonzero value for the government budget.

I

The *MBP* in general suggests that the balance of payments or level of international reserves can be explained via an analysis of the money market. The *equilibrium version* of the monetary approach (call it *MBP*-I) specifies the money market equation in either stock or flow terms and assumes that the money market is in equilibrium. The original Harry Johnson model includes only the money market equation and is of this nature. More complex models of *MBP*-I have been developed which add equations for other markets, but in each model the money market clears.[1] If we assume a money expansion multiplier of unity for simplicity, then money market equilibrium implies

$$(1) \quad M = PL \quad \text{or} \quad \Delta M = \Delta(PL) \quad \text{or}$$
$$\Delta R + \Delta D = \Delta(PL) \quad \text{or}$$
$$\Delta R = \Delta(PL) - 1\Delta D$$

where M = money supply
P = the price level
L = the demand for real money balances
R = international reserves
D = domestic assets of the central bank

Obviously in a small country with no economic growth and with prices and interest rates fixed by foreign magnitudes, we get $\Delta R/\Delta D = -1$. It is important to realize that the ΔR in (1) is the cumulative movement in international reserves and *not* the balance of payments B. This is true because (1) assumes money market equilibrium. Hence B is zero.

The disequilibrium version of the monetary approach (call it *MBP*-II) usually specifies the money market equation in flow terms and considers the money market while it is *out* of equilibrium. There are two

*Associate professor of economics, University of Pittsburgh. Many of the ideas in this paper were developed in conversations with colleagues and students in the Workshop in International Economics at Pitt. Special thanks go to Jim Cassing, Amor Tahari, and Marina Whitman. The work involved here was supported in part by a U.S. State Department Grant No. 5-36579.

[1] See, for example, the long-run solutions in Rudiger Dornbusch and Carlos Rodriguez, as well as the empirical paper by Pentti Kouri and Michael Porter.

variations of this, but they amount to the same thing; one focuses on market excess demands for each good, and the other works through the budget constraint facing the economy. To explain, define three goods: C = commodities, A = bonds, and M = money, and X_i and ϕ_i, where X_C, X_A, X_M = home flow supply of commodities C, interest bearing bonds or financial assets A, and money M minus total (home plus net foreign) flow demand for each good (X_i = home supply of i minus home demand for i minus exports of i plus imports of i) and ϕ_C, ϕ_A, ϕ_M = funds generated from home flow supply of commodities, bonds, and money minus home demand for each good (ϕ_i = home supply of i minus home demand for i).

Each X_i is the excess supply in the home market for good i. Assume no government spending or taxing. The net foreign component of demand is the trade balance (in X_C) and the net capital flow (in X_A). Assume that home money is held only by home residents and that no foreign money is held at home so that there is no net foreign demand term in X_M. Each ϕ_i is the excess supply of good i in the home budget constraint. Confusion arises in the literature because some authors refer to the excess demand for good i as $-X_i$ while other authors use the same term to mean $-\phi_i$.

From my definitions,

$$X_C \equiv \phi_C - BT,\ X_A \equiv \phi_A - KF,\quad X_M \equiv \phi_M \tag{2}$$

where BT = the desired trade balance and KF = the desired net capital flow.

The home budget constraint is

$$\phi_C + \phi_A + \phi_M \equiv 0 \tag{3}$$

From (3) and (2), it necessarily follows that

$$X_C + X_A + X_M \equiv -B \qquad B = BT + KF \tag{4}$$

Frank Hahn and others have used (4) to show that the balance of payments will equal the excess demand ($-X_M$) in the money market if and only if $X_C + X_A = 0$ when transactions occur. Call this MBP-IIa. Ryutaro Komiya, Murray Kemp, John Kyle, and others have explicitly or implicitly *defined* ϕ_C as the trade balance and ϕ_A as net capital flows, so that (3) also gives the balance of payments as identically equal to the home excess demand for money.[2] Call this MBP-IIb.

At first glance MBP-IIb appears to hold always while MBP-IIa requires a special assumption ($X_C + X_A = 0$). However, the assumptions needed for MBP-IIb,

$$[\phi_C \equiv BT \text{ and } \phi_A \equiv KF] \tag{5}$$

can be inserted into (2) to get $X_C \equiv X_A \equiv 0$. If the desired trade balance and net capital flow are identically equal to the excess of home supply over home demand for commodities and assets, then the home commodity and asset markets are *always* in equilibrium; foreign transactions automatically clear each market. The two varieties of the disequilibrium version of the MBP are identical.

When $X_C + X_A = 0$, then (4) can be written

$$X_M \equiv -B \quad \text{or} \quad \Delta D + \Delta R - \Delta(PL) \equiv -B \tag{6}$$

Recall that B is the balance of payments while ΔR is the cumulative movement of R, so that ΔR will always exceed B. Thus, we can write $B \equiv \psi\Delta R$, with $0 \leq \psi < 1$; ψ gradually approaches zero as the money market and the balance of payments adjust toward equilibrium. Consequently, (6) gives

$$\Delta D + \Delta R - \Delta(PL) \equiv -\psi\Delta R \quad \text{or } \Delta R/\Delta D \equiv -1/(1 + \psi) \tag{7}$$

The disequilibrium version of the MBP suggests a small country offset coefficient that is smaller in absolute value than minus unity. Since the money market does not adjust fully, only a fraction of any monetary expansion flows out within the relevant time interval.

II

This section investigates the conditions under which the alternative monetary ap-

[2]The short-run equilibrium analyses in Dornbusch and Rodriguez also utilize such a model.

proaches will give correct conclusions about the level of international reserves or the balance of payments, and how the *MBP*'s relate to more traditional approaches to the balance of payments. From (4) we know that in equilibrium (when all markets clear) each X_i and B are zero. If R and each exogenous variable appear in each X_i and B function then the equilibrium value for R will be the same regardless of which one of the four equations is dropped.[3] Therefore, any model which focuses on the money market (and drops the B equation) should give the same answers as a model which focuses on the balance-of-payments equation. However, *MBP*-I is relevant only if all markets have cleared and the balance of payments is zero. It allows us to tell a story about what has happened to return the money market and the balance of payments to zero.

It is important to understand a crucial difference between the models of the 1970's (which typically have an *MBP*-I flavor to them) and the models of the 1960's. The algebra of more recent models assumes that *all* markets clear when we solve the system for the equilibrium values of all endogenous variables, so that from (4) the balance of payments is always zero in long-run equilibrium. Since the solution value for B (in the long run) is always zero, it necessarily follows that exchange rate or tariff variations cannot alter B. Contrastingly, the models of the 1960's typically require the commodity and money markets to clear, but do not constrain the balance of payments to zero when solutions to the equation systems are found. Thus, in general $B \neq 0$ and it is not surprising that the value for B changes as exogenous variables (such as the exchange rate or tariffs) vary. Moreover, from (4) we know that when $X_C = X_M = 0$, it follows that $-X_A \equiv B$; any balance-of-payments deficit equals the excess demand in the home bond market. Therefore, in the earlier models, $B < 0$ implies $X_A > 0$. The interest rate never rises enough to clear the bond market, and this may have much to do with the different conclusions reached by the two types of models.[4]

From (4) we know that the disequilibrium version of the monetary approach is relevant if the sum of the excess supplies in the commodity and asset markets is zero when transactions occur, $X_C + X_A = 0$. Obviously this will occur if the commodity and asset markets both clear, or if disequilibrium in one is offset by disequilibrium in the other.

There appear to be two different sets of assumptions that will give $X_C = X_A = 0$ (but X_M not necessarily equal to zero) when transactions occur. The first is that the home country be unable to alter the price of any of its home produced goods, which implies that the home country is small and that all of its goods but money are perfect substitutes for goods produced abroad. The perfect substitutes assumption in turn means that there are no nontraded commodities or assets. In such a situation the home commodity and asset markets clear instantaneously, as explained earlier, so that $X_C \equiv X_A \equiv 0$. If the home country is not small or if one or more of its goods are imperfect substitutes for foreign goods, then in general any excess demand or supply will require price and quantity adjustments which take time. Hence, we can have a nonzero X_C or X_A so that $X_M \gtreqless -B$.[5] The state of the money market bears no necessary relationship to the balance of payments.

A second set of assumptions that will insure $X_C = X_A = 0$ and also guarantee that X_M is not necessarily equal to zero any time that $X_C = X_A = 0$ is: 1) there is time for adjustments in interest rates, prices, and other endogenous variables to clear commodity and asset markets; and 2) home and foreign preferences for commodities and assets are identical. The first assumption gives $X_C = X_A = 0$ and is consistent with the contention that the monetary approach is concerned with the long run. The identical

[3] Joanne Salop has proven that this is true.

[4] See Dale Henderson for an excellent review of these differences.

[5] Alan Rabin has concluded something similar for the case when nontraded goods exist at home.

preferences assumption is needed to insure that the money market does not always clear when $X_C = X_A = 0$. With identical preferences there will be a zero net wealth effect from a movement of R from home to foreign residents. Thus, R will not appear in the X_C and X_A functions, and $X_C = X_A = 0$ is possible for all conceivable values of R (only *one* of which will give $X_M = 0$).

The two sets of assumptions which yield $X_C = X_A = 0$, and thereby equate any disequilibrium in the money market with the balance of payments, might be viewed as somewhat restrictive, in that they are unlikely to hold for many or most countries (all of whom have differentiated and/or nontraded commodities and assets), and over relatively brief intervals when all markets are in a state of disequilibrium. However, there are reasonable assumptions which insure that $X_C + X_A \equiv 0$ even when each X_i is nonzero so that *MBP*-II is still relevant.

The requirement is that any disequilibrium in the commodity market creates an equal but opposite disequilibrium in the bond market. One set of assumptions that will achieve this is: 1) any excess demand or supply in the commodity market leads to an unintended change in inventories which must be financed by unintended changes in bond supply; and 2) interest rates adjust quickly so that people are always willing to hold the existing supply of bonds. These assumptions insure that $X_C + X_A \equiv 0$ for all $X_C \neq 0$, because (for example) any excess supply of commodities will create an equal excess demand in the bond market. People willingly hold all bonds, but part of the actual supply of bonds is unintended, so that desired or intended bond supply is less than bond demand.[6]

III

There are two situations where the balance of payments bears no necessary relationship to the money market. First, suppose that the central bank pursues an interest rate target-type of monetary policy rather than using a monetary aggregate as the target. At any point in time the authorities will automatically adjust the money supply so that the money market clears at the desired rate of interest. In this case we have $X_M \equiv 0$ for all values of R and the other endogenous variables, and thus for all values of X_C, X_A and B. Obviously *MBP*-II cannot be used to explain the balance of payments when the authorities control the interest rate.

D. A. Currie has pointed out a second situation where the use of a monetary approach is of questionable significance. This occurs when there is a nonzero value for the government budget. The above analysis and most models in the *MBP* formally ignore government spending and taxing, and the government budget constraint. When we incorporate these, the private sector's budget constraint is

(8) $\phi_C + \phi_A + \phi_M \equiv$ [Income − Taxes − (Demand for Commodities)] + [Private Flow Supply of Bonds − Private Flow Demand for Bonds] + [Flow Supply of Money − Flow Demand for Money] $\equiv 0$

The relationships between X_i and ϕ_i become

$$X_C \equiv \phi_C + T_x - G - BT \quad (9)$$
$$X_A \equiv \phi_A + B^g - KF$$
$$X_M \equiv \phi_M$$

where G = government demand for commodities; T_x = taxes; B^g = flow supply of government bonds.

Combining (8) and (9) gives

$$X_C + X_A + X_M \equiv (T_x + B^g - G) - B \quad (10)$$

If the government finances any budget deficit entirely with bond issues, then (10) reverts back to (4), and all previous conclusions hold. If, however, $B^g + T_x - G$ is nonzero, then in equilibrium (when $X_C = X_A = X_M = 0$) there will be a nonzero value for the balance of payments, which in

[6] Jay Levin has developed a model with these assumptions.

turn must equal that part of the government budget surplus or deficit that is financed by a monetary expansion.[7] The important fact was pointed out long ago by Ronald McKinnon,[8] but until Currie's paper appeared, the profession appears to have been unaware of the effect of all this on some of the conclusions of the monetary approach. First, there is no natural tendency for B to approach zero when all markets (including the money market) clear. Second, if the government budget deficit or surplus is endogenous (as for example when taxes depend on income) then the equilibrium value for B will change when there are shifts in exogenous variables such as tariffs and exchange rates. This, of course, contradicts some of the most important policy implications of *MBP*-I. Finally, when markets are in a state of disequilibrium, the value for X_M will be unrelated to the balance of payments (even if our assumptions keep $X_C = X_A = 0$), so that *MBP*-II is no longer useful.

Some intuitive feel for what is implied by (10) may make all this more palatable. Most *MBP* models contain no government spending or taxing. Hence, from (10) or (4) the sum of the excess supplies in all home markets is equal to the *ex ante* net value of all home private transactions with the rest of the world. When a government sector is included in the model, the sum of the excess supplies in all home markets is still equal to the desired net value of all home private transactions with the rest of the world, but now the latter includes both foreigners and the home government. As William Branson has correctly pointed out, the government budget deficit or surplus plays a role that is identical to the balance of payments. Any private excess demand (supply) for money can be satisfied equally by a government budget deficit (surplus) or payments surplus (deficit).

The limitations to the *MBP* that arise when a government budget is added to the model do not destroy the important fact that the balance of payments is still a monetary phenomenon. A payments deficit can arise when all markets clear, only to the extent that the government is financing at least part of a budget deficit via monetary creation. A budget imbalance that is matched by the sale or repurchase of government bonds will yield a zero balance of payments.

Since $(T_x + B^g - G)$ is equal to the monetary growth rate, the *MBP* is correct in concluding that the balance of payments of a small country will always equal the monetary growth rate. What Currie's work and identity (10) point out is that tariffs and/or exchange rates can conceivably alter $(T_x + B^g - G)$ and turn both the monetary growth rate and the balance of payments into *endogenous* variables. This would be true especially in less developed countries (where the ability to float government bonds is limited) and in countries wherein the authorities seek to peg the interest rate.

[7] Amor Tahari has reached the same conclusions within the context of a growth model for an open economy.

[8] According to McKinnon, "government deficits can be consistent with equilibrium in the private sector of the economy, if at the same time, a trade balance deficit drains off the supply of new financial assets which is being created" (p. 232).

REFERENCES

W. H. Branson, "The Dual Roles of the Government Budget and the Balance of Payments in the Movement from Short-Run to Long-Run Equilibrium," *Quart. J. Econ.*, Aug. 1976, *90*, 345–67.

D. A. Currie, "Some Criticisms of the Monetary Analysis of Balance of Payments Correction," *Econ. J.*, Sept. 1976, *86*, 508–22.

R. Dornbusch, "Currency Depreciation, Hoarding, and Relative Prices," *J. Polit. Econ.*, July/Aug. 1973, *81*, 893–915.

F. Hahn, "The Monetary Approach to the Balance of Payments," *J. Int. Econ.*, Aug. 1977, *7*, 231–49.

D. Henderson, "Modeling the Interdependence of National Money and Capital Markets," *Amer. Econ. Rev.*, Feb. 1977, *67*, 190–99.

Harry G. Johnson, "The Monetary Approach

to Balance of Payments Theory," in his *Further Essays in Monetary Economics*, London 1972.

Murray Kemp, *The Pure Theory of International Trade*, Englewood Cliffs 1964.

R. Komiya, "Economic Growth and the Balance of Payments," *J. Polit. Econ.*, Jan./Feb. 1969, *77*, 25–48.

P. Kouri and M. Porter, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May/June 1974, *82*, 443–67.

John F. Kyle, *The Balance of Payments in a Monetary Economy*, Princeton 1976.

J. H. Levin, "A Dynamic Model of Monetary and Fiscal Policy Under Floating Exchange Rates with Speculative Capital Flows," mimeo., 1977.

R. I. McKinnon, "Portfolio Balance and International Payments Adjustment," in Robert A. Mundell and Alexander Swoboda, eds., *Monetary Problems of the International Economy*, Chicago 1969.

A. A. Rabin, "The Monetary Approach to the Balance of Payments: A Critical Appraisal," mimeo., 1976.

C. A. Rodriguez, "Money and Wealth in an Open Economy Income-Expenditure Model," in Jacob Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, Chicago 1976, 222–36.

J. Salop, "A Note on The Monetary Approach to the Balance of Payments," in P. B. Clark et al., eds., *The Effects of Exchange Rate Adjustments*, Washington 1974.

A. Tahari, "Money, Government Budget Constraint, and Balance of Payments in a Growing Economy," mimeo., 1977.

DISCUSSION

RUDIGER DORNBUSCH, Massachusetts Institute of Technology: John Bilson's paper integrates the main elements in the theory of exchange rate determination and introduces empirical evidence on these questions. The theory emphasizes the role of monetary factors, international linkage of nominal interest rates as well as the relation between interest differentials, and the expectation of exchange rate depreciation. These relations are viewed from the perspective of industralized countries with their efficient highly integrated capital markets.

The main elements of exchange rate determination can be summarized by interest arbitrage, adjusted for exchange rate expectations: $i = i^* + d$ and the definition of the expected rate of depreciation: $d \equiv [\bar{S} - S]/S$, where i and i^* are the domestic and foreign nominal interest rates, and d is the expected rate of depreciation, defined as the percentage excess of the expected future spot rate $\bar{S}$ over the current spot rate S. Combining the two equations yields an expression for the equilibrium spot rate: $S = \bar{S}/[1 + i - i^*]$.

The equilibrium spot rate is equal to the expected future spot rate discounted at the international nominal interest differential. This formulation highlights two essential aspects of exchange rate determination: First, there is the link between exchange rate expectations and the equilibrium value of the current spot rate. Given interest rates, a change in the expected future spot rate is directly translated into an equiproportionate change in the current spot rate. Second, interest rates are a determinant of the spot rate. Given exchange rate expectations, an increase in our interest rate will lead to an appreciation. This establishes the close link between asset markets, in particular short-term money markets, and the exchange rate. The formulation also contributes to an understanding of the volatility of exchange rates which mirrors the volatility of short-term interest rates. That volatility is magnified to the extent that diffuse expectations cause movements in the spot rate to exert a significant effect on exchange rate expectations.

How does such a framework fit the empirical model? Bilson starts from *absolute* purchasing power parity (*PPP*) to derive the exchange rate as the link between national price levels, which in turn are determined by nominal money supplies and real money demands. The equilibrium exchange rate is thus determined in the general equilibrium of the economy, which is proximately characterized by the nominal quantity of money on one side, and interest rates and real income as determinants of money demand on the other side. In a monetarist construction with full price flexibility the theory would suggest that an increase in money would be *instantaneously* offset by an equiproportionate increase in prices and the exchange rate without any incidence on real variables.

Bilson recognizes, of course, that such a view is untenable as a description of the facts. He suggests that *effective* prices are more flexible than measured indices such as the *CPI*. More specifically he argues that the exchange rate can be taken to represent the flexible price segment of the effective price level and deflator of money balances. With such an adjustment an increase in nominal money is dampened in its real effects because the induced depreciation of the exchange rate raises the effective price level and thus reduces the increase in the real money stock. He argues that empirically the exchange rate has a 60 percent weight in the effective price level with the remaining 40 percent going to the *CPI*. Considering that this result arises in the context of monthly data, in equations that use M_2 and M_3 money demands as the aggregates, it vastly overstates the potential effect of the exchange rate. After all, with M_2 and M_3 money demand by the business sector (exposed to the more flexible segment of prices) is deemphasized. For these monetary aggregates we are primarily looking at the behavior of households for whom the *CPI* is, by and large, an accurate measure of the effective price level.

There are more reasons to doubt the exchange rate equation as a correct specification of relative money demands. The derivation assumes absolute *PPP*, an assumption that is at odds with the facts as the work of Irving Kravis and Robert Lipsey and others has shown. Moreover, although different monetary aggregates are used for the two countries (M_2 for Germany and M_3 for the United Kingdom) elasticities are constrained to be equal. Contrary to conventional specifications of money demand there is no allowance for adjustment lags nor for own rates of return in the form of deposit rates. The exchange rate equation should not be interpreted as a test of the money market model but rather as a reduced form that includes the more important macro-economic determinates of exchange rates.

Bilson's interest rate equation is puzzling. It explains changes in the interest differential in terms of regressive and adaptive adjustments to expectational errors. This equation exactly reverses the conventional theory, where short-term interest rates are determined by the balance between money demand and supply and exchange rates adjust to these differentials with minimal if any feedback, to yield the equilibrium term structure of exchange rates. This order of things is indeed supported by the evidence, as can be seen by rewriting the interest rate equation:

$$\underset{}{Dx_t} = \underset{(11.2)}{-.89x_{t-1}} = \underset{(11.4)}{.92\underline{x}_t} - \underset{(.27)}{.03Ds_t}$$

The equation does show a sizeable effect of exchange deterioration on interest rates. A 1 percent depreciation would raise the (annualized) interest differential by thirty-six basis points. It is apparent, however, that the coefficient estimate on the depreciation rate has a large standard error and (assuming zero covariance between the coefficient estimates in Bilson's equation) is not significantly different from zero. In fact the interest rate equation does *not* support a feedback from exchange rates and thus puts in question the dynamic model.

The most striking aspect of Bilson's model is the absence of an explicit role of exchange rate expectations as one of the determinants of the spot rate. It was shown above that short-term interest rate differentials and exchange rate expectations are separate determinants of the spot rate. This is all the more important because, in the well-established absence of *PPP*, nominal interest rates do not contain *all* useful information about the future course of the exchange rate. Exchange rate expectations must be introduced to establish the link between the spot rate and prospective events in the real sector such as commercial policy, fiscal policy, or changes in external indebtedness. They are also important in that they capture the extent to which policy is regarded as accommodating and the paths of money or fiscal policy are viewed as endogenous and responsive to the exchange rate. Failure to include exchange rate expectations in the model is surely one important explanation for the high standard error (5 to 6 percent on a monthly basis) of the exchange rate equation.

JACOB A. FRENKEL, University of Chicago: Peter Kenen's paper contains a useful theoretical framework for the analysis of macro-economic policies in an open economy. Its major characteristics are (i) the detailed specification of portfolio relationships, (ii) the careful distinction between stocks and flows which are linked through the Metzlerian wealth-savings mechanism, (iii) the view that exchange rates are determined by the general equilibrium of the system, and (iv) the assumption that markets clear continuously. Kenen demonstrates that open economies are fundamentally interdependent and that flexible exchange rates do not insulate the economy against all external shocks. I believe that most economists would agree with this conclusion as well as approve of the major characteristics of the conceptual framework. Therefore, rather than discuss the details of the model, I wish to highlight some of the issues and suggest some possible extensions.

The asset-market approach to the exchange rate emphasizes the dependence of the demand for domestic and foreign currency on expectations. Resembling characteristics of other asset markets, current exchange rates incorporate the market participants' expectations concerning future course of events. This perspective is useful in interpreting what otherwise would have been described as "erratic" movements of rates. To incorporate this important feature it would be interesting to relax Kenen's assumption that expectations are "static." Such an extension would incorporate the role of forward markets for foreign exchange as well as the equilibrium condition that is summarized by the interest-parity relationship. It could then be shown that policies which affect the difference between domestic and foreign rates of interest also affect the forward discount on foreign exchange and thereby place an important constraint on the conduct of macro-economic policies aimed at affecting rates of interest.

My second comment relates to Kenen's comparison of the effects that a given policy would have on the equilibrium under fixed and flexible exchange rate regimes. An alternative route would start with the presumption that, independent of the exchange rate regime, the real equilibrium is homogeneous of degree zero in all nominal variables. The analysis would then explore the policies which are likely to be pursued under the two regimes. Thus, while Kenen emphasizes the different effects of a given set of policies, the alternative approach emphasizes the dependence of policy choices on the exchange rate regime.

Kenen's comparison of the two regimes involves the assumption that the various behavioral relationships are given. An interesting extension would relax this assumption. For example, the demand for money under flexible rates might depend on the expected change in the exchange rate reflecting the phenomenon of "currency substitution." To the extent that the structural parameters depend on the exchange rate regime, one may wish to supplement the simulations with an analysis of the stability of the structural parameters as between the two regimes.

Kenen compares the two "clean" regimes. Under the clean float the authorities are presumed not to hold international reserves and not to intervene to affect the rate; monetary policy is assumed to be conducted by open-market operations. Open-market operations amount to an exchange of domestic currency against securities while foreign exchange interventions amount to an exchange of domestic currency against foreign exchange. It is of interest to explore whether there are fundamental differences between these two ways of changing the supply of domestic currency. Furthermore, recent experience reveals that countries have chosen to hold and use international reserves and to manage the float. An interesting extension would examine whether, as an analytical matter, a managed float is just an intermediate system between the two extremes of fixed and floating rates or whether (as I believe) it contains some additional new important elements.

My final remark deals with the wealth effects that are associated with exchange rate changes. Kenen highlights the wealth-savings mechanism and its effect on the trade balance. It might be added that following the capital loss induced by a devaluation, individuals will sell inventories of goods (and bonds) for money in order to restore the desired portfolio composition. This *composition effect* supplements the wealth-savings effect and might dominate the path of the trade balance over the short run.

I wish to conclude with three remarks on Norman Miller's contribution. First, to the extent that economic theory is capable of predicting short-run positions, I would prefer to use the concepts of short-run vs. long-run equilibria rather than those of equilibrium vs. disequilibrium. Second, the phenomenon that when private sector's spending equals its income the deficit in the balance of payments equals the government's budget deficit is an accounting truism that should hold in any sensible

model of the balance of payments (for example, the monetary approach to the balance of payments) since balance-of-payments accounting treats private and government spending alike. Third, to the extent that a change in tariffs induces a finite change in the equilibrium holdings of international reserves (which is the cumulative balance of payments during the adjustment period), the long-run effect on the balance of payments must be zero.

MARC A. MILES, Rutgers University: Marina Whitman originally described this session as an ecumenical session with people of various persuasions. There are papers from people representing three different theologies: an "Orthodox Monetarist," a disciple of the "Church of the Latter Day Keynesians," and a "Unitarian" in the sense that he recognizes the disciples of monetarism to be great philosophers but does not necessarily look to them for salvation. I was probably selected as a token representative of the "Agnostic School," although those of you familiar with what Whitman has called "global monetarism" may think that "World Unification Church" is a more appropriate title. This alternative model has quite different implications than the three papers about the role of exchange rates and foreign account adjustment.

Norman Miller refers to three different types of monetary approach models, one equilibrium and two disequilibrium. His main objection to the equilibrium models is that they fail to explicitly bring in the budget constraint of the government sector. Yet he concludes only that the government deficit will affect the balance of payments to the extent that it is financed by printing money, hardly a new result and one that few monetarists could argue with.

Miller also determines conditions within the disequilibrium models for which "the balance of payments bears no necessary relationship to the money market." Yet all he shows is that in the short run not all the excess supply of money is eliminated, again not a new concept. This intended nonrelationship is an implication of currency substitution models. To the degree that a subset of individuals within countries adjust their diversified currency holdings to changing opportunity costs, flows of money among countries will occur through private markets rather than the balance of payments. Just as Miller wants to show, the domestic money market is no longer reflected in the balance of payments.

The other papers pose two questions about exchange rates. First, what effects do exchange rate changes have? Second, what determines the exchange rate? In an integrated world changes in the relative values of currencies effect only nominal variables. Arbitrage assures that relative price levels do not change, and real variables such as the trade balance remain essentially unaffected. This view contrasts sharply with Peter Kenen's paper. Kenen's model has each country producing only one good. The supply responses of producers in a multigood country that guarantee arbitrage of goods prices are eliminated. If relative price changes occur, producers are not free to increase production of the now relatively cheaper good. The elimination of arbitrage with the assumption of fixed domestic currency prices of produced goods means that devaluation changes relative goods prices. But this relative price change is really a change in the terms of trade, and any resulting effects should be attributed to terms of trade changes, not exchange rate changes.

Kenen also raises the issue of capital gains and losses resulting from the holding of foreign currency denominated securities. A number of papers including my own have analyzed this issue with respect to devaluation and Stephen Magee has done so with respect to currency contracts. While this "bond effect" adds theoretical completeness, it seems to have little empirical importance. Empirical studies have failed to show a net improvement in the trade balance following devaluation. In particular even when I standardized for the effects of government policy and growth, there appeared to be no systematic evidence that

devaluations improve the trade balances. However I have found evidence of some net improvement in the balance of payments. This improvement indicates that exchange rate changes are associated with portfolio readjustments involving the capital account and balance of payments.

Returning to the question What determines the exchange rate?, I would rephrase it to ask: Is the exchange rate determinate? The two papers imply an affirmative answer. Yet from the perspective of the "global monetarist" model a unique exchange rate may not exist. Indeed, there may be an infinite number of possible "equilibrium" exchange rates. The source of indeterminateness is currency substitution in demand as Arthur Laffer and I have argued elsewhere. The more currencies are substitutes in demand, the more the adjustment of the money market will occur through quantity flows rather than price changes. There are still three unknows as shown by Bilson, but now there are only two independent equations.

The degree of substitution should be measured by the elasticity of substitution, not by the interest elasticity as used in Bilson's paper. Therefore Bilson's failure to find an increase in the interest elasticity of the demand for money during the floating rate period has no implication for currency substitution. When I estimated the elasticity of substitution between *U.S.* and Canadian dollars in Canada, I found a significant rise as Canada moved from fixed to floating rates.

Altruism in Law and Economics

By WILLIAM M. LANDES AND RICHARD A. POSNER*

The use of economics to understand the legal system has been growing rapidly. This new field of applied economics is worthwhile for its own sake because the legal system is an important part of the social system. It is also interesting for its potential feedback into the analysis of economic problems in other fields. For example, the analysis of the social costs of crime has led to a change in the thinking of economists about the monopoly problem.[1] Recent work on private law enforcement appears to have broad implications for the problem of employee discipline within a firm.[2] This paper will examine another area where the economic analysis of law appears to have implications for broader economic questions, the law of rescue, and will also explore its relevance to economic questions not limited to the "law and economics" field.

Economists such as Gary Becker have discussed altruism—which we will define as the making of any transfer that is not compensated—mainly in relation to transfers within the family, and secondarily in relation to gifts to charity. Another important area of altruistic activity concerns the rescue of the person or property of strangers. One reads about the passerby who jumps into the lake to save a drowning swimmer—and about the passerby who does nothing to assist the screaming victim of a criminal assault. The question of how to explain either kind of conduct from the standpoint of economics is a challenging one. An examination of the legal regulation of rescue may provide clues to its answer.

I

The peril that invites rescue provides a perfect example of external benefits: *A* sees a flowerpot about to fall on *B*'s (a stranger's) head; if he shouts, *B* will be saved. Thus *A* has it in his power to confer a considerable benefit on *B*. However, it is infeasible for *A* and *B* to contract for the rescue because of the lack of time for negotiation.[3]

A standard reaction to a situation in which there are substantial potential external benefits and high transaction costs is to propose legal intervention. In the example put, this would mean giving *A* the right to either a public or private (i.e., presumably from *B*) reward for the service he renders in saving *B* or punishing *A* if he fails to save *B*. Either form of intervention, however, is apt to be quite costly. Where, as in the example given, the rescuer is not engaged in the business of rescue, the appropriate reward, which from the standpoint of economics depends on the opportunity costs of *A*'s time and the expected costs resulting from the dangerousness to him of the rescue, would be costly to compute. And if the optimal reward was low (because the rescue entailed little cost to *A*), the costs of computation and enforcement of *A*'s legal claim would be high relative to the pure reward component,

*Professor of economics and professor of law, respectively, University of Chicago Law School, and members of the Senior Research Staff, National Bureau of Economic Research. This paper is based on a larger on-going study supported by the Law and Economics Program at the University of Chicago Law School, and by a grant to the National Bureau of Economic Research from the National Science Foundation for research in law and economics.

[1]See Posner and Gordon Tullock.

[2]See Gary Becker and George Stigler.

[3]Actually, the basic cause of the high transaction costs here is not the limited time but, as in more conventional high transaction cost cases, the number of relevant parties: there are simply too many potential rescuers for *B* to identify and negotiate with before he ventures on his walk.

resulting in potentially serious misallocative effects.[4]

The costs of legal intervention are in one important respect reduced under a system of liability for nonrescue (as distinct from a reward for rescue): damages need to be computed only in cases where the rule of liability is violated (or alleged to be violated), and these occasions may be few if compliance with the rule is widespread. The reward approach, in contrast, would require compensation in every case in which a rescue was made successfully. However the liability approach creates another cost: it operates as a tax on activities in which a person may be called upon to attempt a rescue, and like any tax will cause people to substitute away from those activities. This could result in too few potential rescuers, leading to excessive safety precautions by potential rescuees.[5]

The foregoing objections to using the law to internalize the external benefits of a rescue would be much less imposing were it not for altruism, a factor ignored in most discussion of externalities. Altruism may be an inexpensive substitute for costly legal methods of internalizing external benefits—though this depends, of course, on the degree to which altruism will actually motivate rescue.

Becker's analysis of altruistic giving emphasizes wealth disparities between the donor and donee. This emphasis follows from the principle of diminishing marginal rates of substitution, that is, the greater the donor's wealth relative to the donee, the greater the amount the donor is willing to give up at the margin in exchange for a dollar increase in the donee's wealth. The rescue setting presents a dramatic example of wealth disparities. At the moment when the flowerpot is about to crash down on *B*'s head and possibly kill him, *A*, though he presumably does not know what *B*'s wealth was before the flowerpot toppled over, does know that *B*'s expected wealth is now very small and that his own wealth, however slight, is almost certainly much greater than *B*'s. Moreover, if the cost to *A* of effecting the rescue is very small (the cost of a shout), *A* can transfer wealth to *B* at a very low cost to himself. Thus, even though, because they are strangers, *A* presumably values a dollar to himself much more highly than he values a dollar to *B*, the rescue may still be a "profitable" transaction for *A*. Suppose that *A* considers a dollar to be worth a dollar in his own possession but only one cent in *B*'s possession (though if it were not a rescue setting, i.e., if their wealth were more equal, *A* might value a dollar in *B*'s possession at only .01 cent instead of one cent). Nonetheless, if *A* can save *B*'s life at a cost of a dollar and thereby confer a benefit on *B* that *A* can guess is worth several hundred thousand dollars to *B*, the transfer will increase *A*'s utility though he receives no compensation from *B* or anyone else. The "leverage" that *A* obtains by being able to increase *B*'s wealth very greatly at little cost to himself is the counterpart to the matching grant in the conventional charity context, which reduces the cost of a gift to the donor below the dollar amount received by the donee.

The above analysis does not explain, however, why *A* derives *any* utility from the welfare of a complete stranger. This question has generally been elided in economic discussions of altruism. It is assumed that family members (say) have interdependent utility functions but the source of the interdependence is not investigated. Once it is observed that gifts are by no means limited to family members the source of this component of the utility function becomes difficult to accept as a matter of pure assumption.

The biologists have done more work on this question than the economists. They have shown that altruism may increase the likelihood of the altruist's genes surviving

[4]If, for example, the gain to *B* from rescue was $10 and the optimal reward was $1, but the cost of computation $100 and was borne by *B*, *B* might be led to adopt excessively costly safety precautions to avoid being in the position of having to reward *A* for rescuing him. Placing the cost on the taxpayer would have different, but not necessarily less serious, misallocative effects.

[5]For a more complete analysis of the costs and benefits of compensation and liability rules in the rescue setting, see the authors.

in the competition among populations. If insect *A* saves *B* from some peril, this means that *B* will be alive to save *A* should he find himself in danger. Robert Trivers argues this "reciprocal altruism" may enhance the survivorship of the group to which *A* and *B* belong relative to that of some nonaltruistic insect group. Mordecai Kurz has developed an economic model of such behavior. A closely related concept (call it "gene survival") comes into play where, say, *A* in our example dies while saving *B*; *A* and *B* may share some of the same genes and *B*'s survival may contribute more to the chances for the survival of their common genetic endowment than *A*'s.[6]

Reciprocal altruism may explain some, but surely today only a very small, fraction of rescues of strangers. In small communities, the person one rescues, even if a stranger, may indeed be a potential rescuer. In modern urban communities the probability that one is saving someone who will someday reciprocate will often be very close to zero—if he is indeed a stranger. To be sure, the stranger may be carrying some of the rescuer's genes, but this possibility will often be as remote as the possibility that he will someday rescue you. Thus the likelihood that the nonaltruist will be "weeded out" in the competition within or among modern societies is slight.

If we emphasize simply the large discount that the potential rescuer will apply to a stranger's welfare in deciding how much cost to incur in rescuing him, the biological analysis of altruism is helpful. However the analysis seems to imply not only that the discount will be large but that normally it will be so large that only a small fraction of cost-justified rescues (i.e., where the costs to the rescuer are less than the benefits to the victim) would be attempted.

A possible alternative to the biological approach is to emphasize the *recognition* factor in rescues. The fact that most charitable donations are not anonymous and, indeed, that many donors seem quite avid to obtain publicity for their gifts (as where a university chair is named after the donor) suggests that the desire for publicity or recognition is an important factor in charitable giving. Rescuers, too, get their names in the newspapers and this may be the "real" reason why they rescue complete strangers.

This analysis may appear merely to push the inquiry back one step: why do donors, whether of money or services, receive favorable public recognition? Presumably, this results from a public sense, however dim, of altruism as an economizing force (i.e., a low-cost method of internalizing external benefits, compared to legal intervention). Notice that this analysis does not require that *anyone* be in fact altruistic in the sense that he derives utility from making a transfer to a stranger. Conceivably everyone who makes such a transfer does so not out of true altruism but to obtain a reward which consists of favorable publicity.

The importance attached to the recognition factor is relevant to shaping public policy toward rescues. If it is deemed a substantial motivating force in rescues, this would argue against creating liability for failure to rescue. One effect of liability is that the successful rescuer will no longer receive as much favorable public attention, because the public will assume he acted simply out of fear of liability. This increases the tax effect of the liability approach in discouraging potential rescuers.

II

Although the basis for altruistic impulses toward strangers in peril is obscure, the existence of the impulse is verified by the numerous instances in which rescues have occurred where neither reciprocal altruism nor gene survival could provide a plausible motivation. The *fragility* of such impulses—a clear implication of the biological analysis—has also been recognized by the law. Generally the law does not rely on altruism to internalize external benefits where the costs to the rescuer are great. For if the rate at which the potential rescuer equates his costs to the benefits to the

[6]See Richard Dawkins.

person saved is low (for example, it takes $100 in benefits to the person saved to compensate the rescuer for incurring a cost of 1¢), then altruistic rescues are unlikely to occur in cases where the costs of rescue are large.

Two examples will illustrate the law's recognition of this point. Although the ordinary rescuer is entitled to no reward, the professional (normally a physician) is entitled to collect his standard fee from the person rescued in the high-transaction-costs setting (for example, no negotiation is possible because the victim is unconscious). Not only is the physician's opportunity cost of time higher than that of the average nonprofessional rescuer, but, because of his greater knowledge of medical treatment, he is expected to spend more time with the rescued person (treating him, as distinct from simply calling an ambulance). Thus the total costs of rescue to the physician are apt to be higher than those borne by the nonprofessional. (To some extent, however, the greater benefit normally conferred by the professional rescuer's more extensive services may offset the added cost.) The costs of computing the reward, moreover, are relatively slight because the physician's fees for similar services are readily discoverable.

The second example is rescue at sea. Normally this is undertaken by commercial operators (especially in cases where the vessel or its cargo, rather than just passengers and crew, are salvaged) at substantial cost. One is not surprised that a successful rescue at sea entitles the rescuer to a reward—and that the rescuer's right is most firmly established where it is property rather than lives that is rescued: the cost of pure life salvage is normally less than that of property salvage, and the normally greater value of lives versus property increases the likelihood of an altruistically motivated rescue of lives.[7] An additional factor is that to the extent rescue is undertaken by firms operating in a competitive market, as is usually the case at sea, the costs of altruism to the rescuer tend to be very great; the firm's very survival may be at stake because altruism implies the bearing of uncompensated costs that a nonaltruistic competitor would avoid. A closely related point is that altruism is not a trait with positive survival value in a competitive market. On the contrary, competition will tend to weed out the altruistic seller, just as it tends to weed out any other type of high cost seller.

Given that legal intervention and altruism are substitute methods of encouraging the internalization of the external benefits of rescues in emergency situations, the question naturally arises whether studying the pattern of legal intervention in rescues might provide a clue to variations over time or across societies in the level of altruism. We have compiled a list (available on request) of the countries (and a single *U.S.* state—Vermont) that impose liability for failure to rescue, by date of first imposition of liability. The task of explaining this ordering is a formidable one and we are not able to offer more than conjecture. It may, however, be significant that no law imposing liability for nonrescue has been found prior to 1867. This may reflect the fact that in a preurban society reciprocal altruism may provide an adequate substitute for legal coercion to rescue.

Another suggestive feature of our list is the predominance of fascist and communist states among the early adopters of liability for nonrescue. Liability for failure to rescue is a form of conscription for social service which would seem congenial to a state that already regards its citizens' time as public rather than private property. It may not be accidental that the first (and thus far only) *U.S.* state to impost liability for nonrescue is Vermont, which has the third highest tax rate (after Alaska and New York) in the United States.

III

Thus far we have discussed altruism as a substitute for law in internalizing external benefits. Why should it not equally be a

[7]For a detailed discussion of professional rescue and some empirical analysis of salvage awards, see the authors.

substitute for law in internalizing external *costs*? Indeed, if we do not need a law to compel rescues, why do we need, for example, a law to compel drivers to avoid running down pedestrians?

The reason would appear to lie in the significant discount the driver is likely to attach to pedestrians' benefits and the high cost of accident avoidance (for example, damage to one's car and personal injury, or the cost of altering one's behavior at an earlier stage, such as driving at a slower speed, to avoid situations in which an accident is imminent). To be sure, when these costs are low, even a relatively small degree of altruism will be sufficient to induce the driver to avoid the accident. When these costs are substantial, though not as large as the benefits to the pedestrian, altruism is unlikely to be an adequate method of internalizing the pedestrian's losses and hence a liability rule will be required to generate optimal accident avoidance.

Why, therefore, does not society impose liability *only* when the costs of avoidance are high (though still less than the victim's benefits) and rely on altruism alone to prevent low-avoidance-cost accidents? This approach would be symmetrical to the treatment of compensation in the rescue setting. However, the principal objections to compensation in the low-cost rescue case—the cost of computing the reward, the cost of transacting between the parties, and the possible use of costly legal proceedings to enforce one's right to a reward—are not present when the question is whether to impose liability in the low-avoidance-cost accident situation. Here a liability rule, if effective, will be relatively cheap because the accident will usually be deterred.

REFERENCES

G. S. Becker, "A Theory of Social Interactions," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1063–93.

——— **and George J. Stigler**, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *J. Legal Stud.*, Jan. 1974, *3*, 1–18.

Richard Dawkins, *The Selfish Gene*, New York 1976.

M. Kurz, "Altruistic Equilibrium," Inst. for Math. Stud. in Soc. Sci., Econ. Ser., rept. no. 156, Stanford Univ. 1975.

W. M. Landes and R. A. Posner, "Salvors, Finders, Good Samaritans, and Other Rescuers: An Economic Study of Law and Altruism," *J. Legal Stud.*, Jan. 1978, 7.

R. A. Posner, "The Social Costs of Monopoly and Regulation," *J. Polit. Econ.*, Aug. 1975, *83*, 807–27.

R. L. Trivers, "The Evolution of Reciprocal Altruism," *Quart. Rev. Biology*, Mar. 1971, *46*, 35–56.

G. Tullock, "The Welfare Costs of Tariffs, Monopolies, and Theft," *Western Econ. J.*, June 1967, *5*, 224–34.

Capital Punishment and Homicide in England: A Summary of Results

By KENNETH I. WOLPIN*

This paper is a summary of a more extensive examination of the time-series pattern of homicides in England and Wales from 1929 to 1968.[1] The analysis is conducted within the structure of an economic approach to criminal behavior, that is, the aggregate national observations are hypothesized to have been generated by the cumulation of individual decisions based upon a maximizing calculus such as that of Isaac Ehrlich among others. Whether such a behavioral model is applicable to the crime of homicide is the major empirical issue addressed in this essay. Although many homicides result from seemingly little prior reflection of costs and benefits, precipitating circumstances are themselves not necessarily beyond choice. The degree to which the expected penalty for homicide influences the decision of avoiding disputes leading to violence cannot merely be presumed negligible.

The economic framework asserts that individuals respond to the structure of incentives and penalties embodied in the criminal justice system. A direct implication is that increases in expected punishment through changes in probabilities of arrest and conviction will reduce the incentive to commit homicide. Moreover, among those actions which retard criminal behavior, effectiveness, as measured by the offender's elasticities of response, can be ordered; a given percentage rise in the probability of arrest is predicted to reduce the homicide rate by a greater proportion than an equal percentage increase in the conditional (upon arrest) probability of conviction and the latter by a greater proportion than the same percentage increase in the conditional (upon conviction) probability of execution or imprisonment. The intuition is that increases in the likelihood of events which are conditional upon fewer previous events place an individual in jeopardy of entering a greater number of further undesirable states, that is, states in which additional costs are incurred.

The theory also implies that increases in the severity of punishment diminish homicidal behavior. If execution is the most severe punishment, it will be the greatest deterrent. In principle, the question of whether homicides conform to the economic perspective can be divorced from the question of whether executions are a more effective deterrent than alternative forms of punishment as practiced, though the question of the deterrent value of any specific form of punishment is fundamentally inseparable. Acceptance of a general theory of deterrence does not require acceptance of a *differential* deterrent impact of capital punishment.

The evidence from the English experience lends support to the deterrence perspective and to the notion that execution is viewed by potential offenders as more severe treatment than the level of the alternative imprisonment penalty. Regression estimates indicate a differential negative impact of executions on homicides over that of "life" imprisonment that is roughly stable over different specifications and time periods. Several more casual pieces of evidence also conform to the same conclusion. The fact that there are im-

*Assistant professor, department of economics and Institution for Social and Policy Studies, Yale University. This paper was prepared under Grant Number 75NI-99-0127 from the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice. Points of view or opinions stated in this document are my own and do not necessarily represent the official position or policies of the U.S. Department of Justice. I wish to thank Randall Olsen for many useful discussions recognizing the usual disclaimer of responsibility for remaining errors.

[1]The more detailed paper is available from the author on request.

portant limitations of the data and difficult methodological issues yet to be resolved should serve to caution the reader against forming policy implications from any particular point estimate even if it passes "conventional" tests of significance. From a policy perspective, the appropriate null hypothesis need not correspond to the usual one of zero effect if there exists a positive probability of conviction and execution of "innocent" individuals. In addition, the *ceteris paribus* conditions inherent in the regression estimate need not be appropriate in making a policy assessment.

I. The Data

National statistics on homicides for England and Wales, divided into murders and manslaughters, have been annually published in the same basic format since 1894. As of 1929, a more detailed set of statistics are available for homicides involving victims over (and under) a year old which ultimately (i.e., through court proceedings) were classified as murder. The primary difference between the two series for murder is that for the shorter period the cases in any particular year are followed to termination through the following year. The individuals who are arrested in any year correspond exactly to those convicted and punished even if the later procedural stages occur in the following year. The data used in this analysis combine the murder statistics of the shorter period with the murder and manslaughter statistics from the longer period, both exclusive of victims under a year old, to form a homicide series for the 1929–68 period which tracks individuals as well as possible through the different criminal justice stages. A detailed description of the information contained in the published statistics is provided in Figure 1. The chart is essentially self-explanatory and only a few remarks are necessary.

Upon the commission of a homicide, either an arrest is made for murder or for manslaughter, or the offender is not captured. In more than a few cases, the suspected offender commits suicide prior to arrest and these crimes are considered to be cleared in the same way as if an arrest had been made. If the individual is tried for manslaughter, either a conviction or acquittal ensues. Following a murder arrest, an individual may be convicted or acquitted of murder, found to be "guilty but insane" or

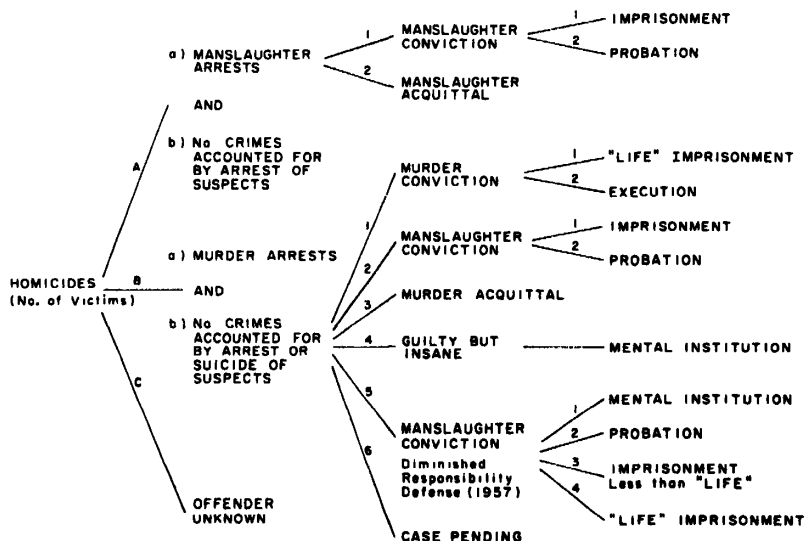



Figure 1

convicted of manslaughter. The finding guilty but insane corresponds to the notion of not guilty by reason of insanity. Until 1957, it was an often observed trial verdict.

In 1957, the Homicide Act greatly affected the use of this verdict and made several more profound changes in the law as well. The most important modification was the creation of classes of murder identified as capital (i.e., subject to execution) as distinct from those treated as noncapital. The major types among those that were capital were murders committed in the furtherance of theft, murders by shooting or explosive device, and murders of police officers and prison officers. Prior to 1957 all convicted murderers were subject to execution. In fact, the death sentence was mandatory given conviction for murder. However, it was not always carried out. Upon the advice of the Home Secretary of State, a Royal Prerogative of Mercy could be issued commuting the death sentence to life imprisonment. From 1900 to 1955 a little more than one-half of all convicted murderers were executed. That the 1957 Homicide Act was an abrupt policy change is unquestionable; in the five years subsequent to the Act only 10 percent of convicted murderers were executed while in the five years preceding the change in law from 1951–55 (in 1956 there were no executions in anticipation of the legal change), 52 percent of those convicted of murder were executed. In absolute terms, in the five years prior to the change 73 executions occurred while in the five years after there were 23.

The other important change embodied in the Homicide Act was the creation of the defense of "diminished responsibility," whereby sufficient evidence of impairment of a defendant's mental responsibility at the time of the act could reduce the conviction to manslaughter. One effect of this change was to greatly reduce the number of verdicts of guilty but insane. It is not evident that all manslaughter convictions arising from this defense would previously have entailed a murder finding of insanity. Several conventions could have been employed to classify these individuals. The one reported here was to assign to the guilty but insane category of murder only those persons who were sentenced to mental institutions. Those sentenced to life imprisonment were assumed to be murder convictions while those given a probationary sentence or imprisoned for less than life were assumed to correspond to previous manslaughter convictions.

Two other legislative changes affected the published statistics, the Murder Act of 1965, which abolished the death penalty and substituted a mandatory life imprisonment term and the Road Traffic Act of 1956, which created a new category of manslaughter classified and reported separately as deaths caused by dangerous driving. Manslaughters caused by dangerous driving were combined with other deaths caused by dangerous driving making it impossible to add back to manslaughters those deaths that would have been classified as manslaughters before the Act. The major effect was a drop in the number of reported manslaughters and arrests for manslaughter primarily since 1957, as the Act was instituted in November of 1956. Available evidence suggests that the number of manslaughter convictions was only negligibly affected as a large number of such arrests resulted in acquittal. Arrests for manslaughter fell from 105 in 1956 to 61 in 1957 while manslaughter convictions remained roughly constant.

II. Variable Definitions, Empirical Specifications, and Regression Results

Table 1 defines, in terms of the classification scheme of Figure 1, all of the crime control variables used in the analysis as well as environmental variables reflecting demographic and economic conditions. The general specification of the homicide supply equation includes a measure of the clearance rate, conviction rate upon arrest, and type of punishment upon conviction. Since the proportion of homicides cleared either as murder or manslaughter was roughly constant over the period, its mean being .963 and its standard deviation .019, it

TABLE 1—ORDINARY LEAST SQUARES HOMICIDE SUPPLY REGRESSIONS[a]
(t-statistics in parentheses)

	1 1929-68	2 1929-68	3 1929-68	4 1929-68	5 1929-55
MDCLRAT: $\frac{Bb}{\text{Homicides}}$	−.576 (3.99)	−.587 (4.42)	−.773 (5.29)	−.571 (4.17)	−.502 (2.18)
CNMDAMD: $\frac{Ba1 + Ba54}{Ba - Ba6}$	−.0005 (.005)	−.017 (.19)	−.038 (.34)	−.005 (.06)	−.036 (.25)
CNMNAMN: $\frac{Aa1}{Aa}$	−.087 (1.47)	−.105 (1.87)	−.132 (1.97)	−.104 (1.84)	−.187 (1.62)
GTINAMD: $\frac{Ba4 + Ba51}{Ba - Ba6}$	−.171 (2.87)	−.136 (3.17)	−.073 (1.55)	−.135 (3.09)	−.142 (1.64)
IMPCNMN: $\frac{Aa11 + Ba21 + Ba53}{Aa1 + Ba2 + Ba52 + Ba53}$	−.186 (1.14)	−.190 (1.25)	−.245 (1.33)	−.177 (1.14)	−.335 (.76)
ECXNMD[b]: $\frac{Ba12}{Ba1 + Ba54}$	−.189 (3.32)	−.140 (4.66)	—	—	−.115 (1.08)
*ECXNMD*B56*	—	—	−.190 (2.36)	−.099 (1.36)	—
*EXCNMD*A56*	—	—	−.071 (2.55)	−.148 (4.45)	—
UNEMPL	.248 (2.61)	.209 (3.60)	.218 (2.96)	.197 (3.19)	.275 (1.84)
GDPPOP	.137 (.23)	.359 (1.96)	.459 (2.00)	.319 (1.62)	.370 (.89)
M2029M	−.450 (.63)	—	—	—	—
URBPOP	−1.721 (1.15)	—	—	—	—
DUMY57	−.515 (2.60)	−.359 (3.80)	—	−.411 (3.25)	—
AW	.503 (2.81)	.408 (3.29)	.468 (3.09)	.391 (3.04)	.506 (2.31)
YR	−.002 (.11)	—	—	—	—
C	−15.281 (5.18)	−14.881 (18.43)	−15.513 (14.86)	−14.629 (16.01)	−15.036 (7.30)
R^2	.836	.819	.733	.822	.642
D. W.	2.048	1.888	1.863	1.921	1.860
Number of Observations	33	33	33	33	20

[a]All variables are in natural logarithms. The dependent variable is homicide per capita. Variable definitions are given according to the legend in Figure 1 except for the following which are defined as follows: *M2029M*: the proportion of all males within the ages of 20-29; *UNEMPL*: the *U.K.* unemployment rate net of temporary layoffs; *URBPOP*: the proportion of the population residing in nonrural areas; *GDPPOP*: the real gross domestic product per capita for the United Kingdom; *AW*: a dummy variable with value 0 before World War II and value 1 after; *DUMY57*: a dummy variable with value 0 before 1957 and value 1 from 1957 to 1968; *YR*: a continuous time trend taking the values 1-33 and corresponding to the years 1929-68 exclusive of World War II years.

[b]In 1956, 1966-68 there were zero executions. It was assumed that there was 1 execution in forming the logarithm of the execution rate for those years.

was impossible to ascertain a precise estimate of its effect on the homicide rate. Instead, a murder-specific clearance rate, that is, the proportion of homicides cleared as murder, was substituted. Given the available data the conviction rate was disaggregated into the proportion of all murder arrests leading to murder convictions (*CNMDAMD*) and the proportion of all manslaughter arrests leading to man-

slaughter convictions (*CNMNAMN*). Since the guilty but insane category applies only to murder, its probabilistic representation (*GTINAMD*) uses murder arrests as the denominator. Punishment variables are also specific to the type of conviction-execution taken relative to murder convictions and manslaughter imprisonment relative to murder convictions.

Table 1 reports the regression results. The regression equations were estimated in linear and logarithmic (*log-log*) forms, though only the latter is reported here. Results are essentially identical. Serial correlation of the usual first-order variety is not present so only ordinary least squares regressions are reported. Discussion of simultaneous equations issues are contained in the more detailed paper. The inclusion of *DUMY57* is intended to reflect the severe drop in reported manslaughters due to the Road Traffic Act. The dichotomous variable distinguishing the post-World War II period (*AW*) is inserted to capture any permanent influences on behavior associated with the war.

The impact of each of the crime control variables is estimated relative to some alternative, the complement to the murder clearance rate being the proportion of homicides not cleared or cleared as manslaughter, that of the murder conviction rate being the proportion of murder arrests in which the defendant was either acquitted of murder or manslaughter or convicted of manslaughter, and that of the manslaughter conviction rate being the manslaughter acquittal rate. The alternative punishment for manslaughter is probation while the alternative punishment for murder is life imprisonment. Of course, life imprisonment does not literally imply incarceration until death, for the vast majority of convicted murderers are released before that event. Unfortunately published figures do not exist for this period on the average time served for convicted murderers, an omission of potential importance discussed in my detailed paper.

Increasing the proportion of homicides cleared as murder with a concomitant decline in homicides cleared as manslaughter (or unsolved) reduces the number of homicides. Moreover, the elasticity of response to such a change far exceeds that of any other "deterrent" variable, as was hypothesized in the economic model. However, elasticities of response to conviction rate changes generally fall below responses to punishment rates, though the imprecision of several of these estimates prevents firm conclusions.

Before turning to the estimates of execution effects, it should be noted that the environmental variables *M2029M*, *URBPOP*, and *YR* do not jointly affect the homicide rate or the crime control parameter estimates. That the age distribution variable in particular does not explain much of the rise in the homicide rate since 1957 is not surprising given that the fraction of males aged 20 to 29 did not begin to increase noticeably until around 1963 or 1964.

The point estimate on the proportion of convicted murderers actually executed (see equation 2) implies that, with no adaptations by the police, juries, or potential offenders which alter other deterrent variables, executing an additional convicted murderer evaluated at the average execution rate, the average number of convicted murderers, and the average murder rate for the entire sample, would reduce the number of homicides by .0932 per million population, which, evaluated at the average population, comes to 4.08 potential victims. Note that this effect is relative to the alternative punishment of an imprisonment sentence which does not generally exceed 10 to 15 years and that the standard deviation of that estimate is .88 victims.

If it were not for lack of space, many more pages would follow critically assessing the evidence briefly presented above. In my more detailed paper, I offer explanations of changes in the homicide rate other than deterrence. With some, available information seems insufficient to provide clear empirical tests of the alternatives. The deterrence hypothesis remains, however, a view that is on balance consistent with the English data on homicides.

REFERENCES

I. Ehrlich, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *Amer. Econ. Rev.*, June 1975, *65*, 397–417.

K. Wolpin, "Capital Punishment and Homicide: The English Experience," unpublished paper, Yale Univ., Oct. 1977.

Great Britain Central Statistical Office, *Annual Abstract of Statistics*, 1894–1967.

Great Britain Home Office, *Judicial Statistics, Parlimentary Papers of the House of Commons*, 1894–1967.

The Subtle Impact of Price Controls on Domestic Oil Production

By RODNEY T. SMITH AND CHARLES E. PHELPS*

Price controls generally are believed to reduce *U.S.* oil production. Allegedly, the production disincentive equals the foreign controlled domestic price disparity, which has fluctuated between 50 to 60 percent of the world oil price during 1974–75. (See Robert Hall and Robert Pindyck.)

The effects have been much different. Initially, the actual control program had an ambiguous impact on the level—but increased the decay rate—of domestic oil production. The effect of the control program becomes an empirical issue and preliminary analysis of 1974–76 data shows only a small impact. However, the controls had a provocative impact on the elasticity of domestic oil production with respect to the world price. With the administrative changes accompanying the 1975 Energy Production and Conservation Act (*EPCA*), theory implies that the elasticity became negative.

I. The Oil Price Controls

The oil price controls stressed the balancing of two conflicting goals: controlling capital gains oil producers enjoyed from increasing oil prices and providing incentives for increased oil production. To these ends, the government attempted to place price controls on inframarginal oil while allowing marginal oil production to be sold at uncontrolled prices. Administrators faced the Gordian task of determining when oil was inframarginal. Rules of thumb were developed and often changed.

Two primary categories of oil were originally created: "old oil" and "new oil." (Production from "stripper" wells producing less than ten barrels per day was exempted from controls.) Old oil referred to monthly production from properties in the base year 1972. The old oil price initially could not exceed the May 15, 1973 posted price, which was less than half of uncontrolled prices after world oil prices increased in late 1973. New oil included all output exceeding base (1972) production. New oil could be sold at uncontrolled prices. This scheme was enforced by the old oil allocation program, which assigned legal rights to controlled oil according to contracts in force in December 1973.

The rules were immediately amended to increase domestic oil production. The price ceiling on old oil was increased by $1.35 per barrel. Also for each barrel of new oil the producer could "release" a barrel of old oil from under the price controls. That "released oil" could be sold at uncontrolled prices, with measured old oil falling by those quantities.

To eliminate effects of differences among refiners' access to controlled oil, the entitlement program was instituted in November 1974. That program implemented a rationing ticket scheme for controlled oil. The refining of a barrel of controlled oil was accompanied by the liability of obtaining a fraction of an additional entitlement ticket, thus reducing the recipient's marginal value of controlled oil below the uncontrolled price by about $2.50 to $3.00 per barrel.[1]

The 1975 *EPCA* modified the controls. The released oil provision was abolished. Old and new oil were redefined to be "lower" and "upper tier" oil, respectively. The base year for defining upper tier oil changed from 1972 to 1975 for properties producing in 1972. For properties produc-

*Economist, The Rand Corporation, and visiting assistant professor, Claremont Mens College; and senior economist, The Rand Corporation, respectively.

[1]For vertically integrated producer, the entitlement liability essentially levied a tax on production equal to the difference between the value of uncontrolled oil and the price ceiling.

ing only old oil during 1973–75, base level production in subsequent years was to be reduced annually at the same property-specific rate that production fell during 1973–75. Upper tier oil was brought under price controls, although the allowed price was above lower tier oil. Stripper oil was similarly treated for seven months, then decontrolled. All price ceilings had general inflation adjustment clauses and the ceilings were scheduled to be removed in April 1979.

II. Impact on Domestic Oil Production

This discussion focuses on the marginal revenue from domestic oil production and the user cost of oil with and without price controls. Throughout we assume that the controls had no effect on the world price of oil. Two considerations dominate the analysis. First, the frequent changes in petroleum regulation generate many different regulatory regimes since 1973. Major changes include the entitlement program, the supplemental tariff on oil, and the 1975 *EPCA*. Second, the controls differentially affected independent versus integrated producers and production from "old" properties existing before 1972 and from "new" properties entering after 1972.

A. *The Independent Producer on Old Properties*

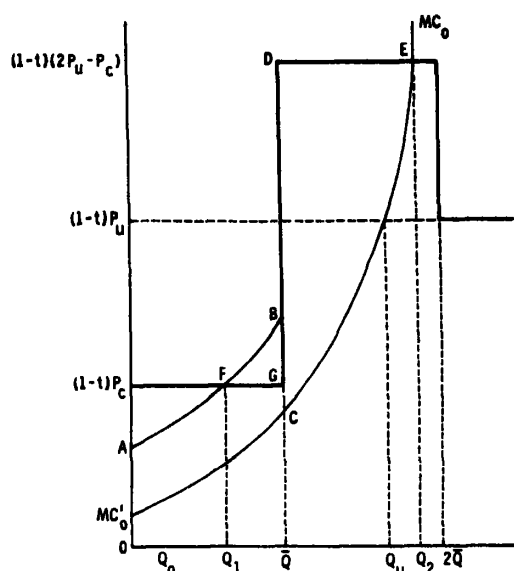
In the absence of controls (see Figure 1), oil producers would receive for each barrel produced[2] the uncontrolled price P_u less *ad valorem* federal royalties and severance taxes tP_u of at least 20 percent (see Richard Mancke). During initial controls, the price line became discontinuous. For rates of production (Q) less than base period production ($\bar{Q}$), the producer's after-tax and royalty marginal revenue was $(1-t)P_c$. (Since P_c was increased by $1.35 after

Base production $= \bar{Q}$
New oil $= Q_2 - \bar{Q}$
Released oil $= \bar{Q} - Q_0 = Q_2 - \bar{Q}$
Old oil $= \bar{Q}$ – Released oil $= 2\bar{Q} - Q_2 = Q_0$

FIGURE 1. THE INDEPENDENT PRODUCER UNDER PRICE CONTROLS

the controls' inception, base production $\bar{Q}$ lies to the left of the intersection of the MC_0 curve with the $(1-t)P_c$ line.) For $\bar{Q} < Q \leq 2\bar{Q}$, the released oil provision increased the marginal revenue to the uncontrolled price P_u, plus the bonus differential $(P_u - P_c)$ on a released old oil barrel, giving after-tax/royalty marginal revenue of $(1-t)(2P_u - P_c)$. For $Q > 2\bar{Q}$, the after-tax/royalty marginal revenue equalled $(1-t)P_u$.

Marginal extraction costs were also affected by the control scheme. Without controls, the marginal cost curve would reflect current marginal costs of extraction and the user cost of oil associated with anticipated changes in future oil prices versus future marginal costs of extraction. For convenience, we assume that any nongovernment increase in the current uncontrolled oil price is expected to continue into the future. These marginal cost conditions are reflected in the curve MC_0 and are assumed to be unaffected by the controls. With the

[2] The reported uncontrolled price of domestic production need not be identical to the reported foreign oil price because of quality differentials and transfer pricing by multinational firms. In 1975, *FEA* data showed the former less than the latter by about $1.50 per barrel.

controls a user cost arises due to anticipated decontrol. Oil sold currently under price controls can be held underground until controls are repealed. The magnitude of this user cost depends upon the expected date of decontrol, the resulting price increase, and the interest rate. Reflecting this user cost, the marginal extraction cost curve under controls becomes the discontinuous curve $ABCMC_0$.

Controls affect both initial production levels and rates of decay. For the situation in Figure 1, two candidates Q_1 and Q_2 occur for initial production under controls, which straddle uncontrolled production Q_u. Level Q_2 is optimal if returns from extending production from $\overline{Q}$ to Q_2 (area CDE) exceed losses from extending production from Q_1 to $\overline{Q}$ (the area BFG). If so, the property would have new and released oil production of $Q_2 - \overline{Q}$ each, the remainder being old oil production $(2\overline{Q} - Q_2) = Q_0$. Alternatively, if the lower production rate were optimal, then only old oil production of Q_1 would arise. Thus the released oil provision of price controls may initially have increased *U.S.* oil production.

Over time the released oil provision's effect erodes because of reserve depletion. Depletion shifts to the left the entire marginal extraction cost curve. The two potential production levels Q_1 and Q_2, and production under no controls Q_u, all decline. Eventually the lower production rate becomes optimal because the area BFG grows and the area CDE shrinks. Consequently, the production decay rate increases as properties transfer from producing new and released oil to producing only old oil. Also, the increased user cost of oil due to regulation (the segment AB in Figure 1) increases the speed at which fields are withdrawn from production.

Another possible situation occurs when the marginal extraction cost intersects the marginal revenue curve at a production rate exceeding $2\overline{Q}$. Production under price controls would equal production without controls. As depletion occurs and production enters the range $\overline{Q}$ to $2\overline{Q}$, the previous analysis applies.

For *U.S.* oil production during the controls, total costs were not minimized. Properties with released oil had marginal costs well in excess of the uncontrolled price, while those producing only old oil had marginal costs substantially below the uncontrolled price.

The situation remained unchanged until imposition of the supplemental crude oil tariff in 1975. Initially \$1.00/barrel in February 1975, the fee increased to \$2.00/barrel by May and resulted in increased uncontrolled oil prices until its removal in January 1976. (The analysis is identical for any change in uncontrolled oil prices.) For properties producing new and released oil marginal revenue changes were twice the change in price, net of tax/royalty. The increased uncontrolled price also raised the user cost of oil due to anticipated deregulation whenever the tariff was anticipated to be sustained, possibly affecting which equilibrium production level was selected. However, this increased user cost was less than the increase in the $P_u(1 - t)$ price because of discounting and any belief that the tariff would be repealed before deregulation. On balance, production on such "releasing" properties would have increased due to the crude oil tariff. Production increases would exceed increases forthcoming under no price controls.

For properties producing only old oil, the tariff introduced two possibilities. First, production of new oil could become profitable, so that oil production would increase and be above the level forthcoming without controls. Alternatively, production of new oil could remain unprofitable. However, old oil production would still decline in response to the increased user cost of oil. For these properties, rates of production became *inversely* related to the uncontrolled price! This result is significant since up to four times as many properties produced only old oil in early 1975 as produced some released oil (see the authors).

As indicated above, the 1975 *EPCA* changed the control environment and repealed the supplemental oil tariff. Consider a property producing only old oil in 1975. Base period production was

redefined to equal 1975 production, so that marginal revenue increased in the neighborhood of 1975 production from the lower to the upper tier price. While production would initially increase, these properties will eventually cease producing upper tier oil as depletion occurs. For a property producing new oil in 1975, the *EPCA* reduced marginal revenue in the neighborhood of 1975 production as the released oil provision was abolished and new oil prices fell to controlled upper tier prices. Oil production on those properties declined. Initially, production stimulated by redefinition of the base period probably dominated the production sacrificed by abolishing released oil, since some 70 percent of all properties produced only old oil. However, production increases must be short-lived, as depletion erodes these gains.

The 1975 *EPCA* undoubtedly generated an *inverse relationship* between domestic production and the world price of oil. By bringing all properties under price controls, increases in the real price of world oil raise user costs of oil associated with anticipated deregulation for all properties.

B. *The Integrated Producer-Refiner on Old Properties*

Oil production incentives were initially different for vertically integrated firms. We analyze the fully integrated producer, and then generalize to the partly integrated producer, the predominant form of organization.

Under initial controls, fully integrated producers would ignore crude oil price controls when choosing oil production levels. Producers valued at the uncontrolled price P_u any controlled oil produced for an affiliated refiner, since that oil substituted for oil purchased at uncontrolled prices. While legal price ceilings for refined products declined with the use of old oil, those ceilings were generally nonbinding and thereby would not affect oil production decisions (see the authors).

Controls conferred a benefit to these firms. Reduced payments of royalties and severance taxes resulted because those payments were based on the controlled price P_c, rather than P_u. In effect, controls legalized advantageous transfer pricing. While these firms had incentives to transfer price prior to controls, governments and owners of mineral rights had incentives to restrain this activity.

For integrated firms, oil production was unambiguously increased by initial controls. Marginal revenue under controls exceeded marginal revenue without controls by about 15 percent, because of reduced royalty and severance tax payments. Marginal extraction costs were unaffected by controls, reflecting the irrelevance of price ceilings to such firms.

The entitlement program significantly altered this environment. Production and use of controlled oil became accompanied by an entitlement liability, a tax equal to the difference between uncontrolled and controlled oil prices. Fully integrated producers now faced the same production incentives and disincentives as independent producers under the initial control scheme, except that the entitlement program reduced the value of uncontrolled oil below P_u for vertically integrated producers.

In conceptual terms, placing fully integrated firms under controls via the entitlement program involved the same forces as placing independent producers under controls in 1973. Thus production by integrated firms could have increased or decreased. The analysis of the crude oil tariff and the *EPCA* follows the earlier analysis of these events for independent producers.

For partly integrated producers, incentives are a mixture of those for integrated and independent producers. The Allocation Program required controlled oil prorationing among buyers according to December 1, 1973 contracted quantities. Thus, production incentives and disincentives for partly integrated producers are a weighted average of those prevailing for independent and fully integrated producers.

C. *Production on New Properties*

New properties reflect both exploration and extraction activity for long-run expan-

sion in production, but incentives for exploration undoubtedly dominate long-run effects of the controls. Prior to the 1975 *EPCA*, production on all new properties could be sold at uncontrolled prices, so controls had no direct effect on production or exploration. If regulatory changes were anticipated, effects on exploration and production would follow usual analyses of anticipated regulatory changes. After the 1975 *EPCA*, exploration and production from new properties would decline, as indicated by standard analysis of price controls.

III. Empirical Analysis

Assessing the quantitative significance of the controls requires property-specific data on oil production. Since such data are unavailable, we discuss aggregate *U.S.* production data and must accept the problem from ignoring differences among effective control schemes.

Any examination of recent data must consider important changes in economic and legal environments besides the controls. Quadrupling world oil prices would undoubtedly increase oil production without controls, although the total production response would not be immediate. The 1970's have also witnessed abandonment of state government prorationing and abolition of the depletion tax allowances for most oil producers. Repeal of prorationing, predating 1974, would increase oil production once and for all and thereby confound the use of pre-1974 data as information on oil production trends without controls. Fortunately, the depletion allowance probably did not significantly affect domestic oil production (see Mancke) so that its repeal does not confound the analysis.

Table 1 shows our estimated regression equation explaining monthly *U.S.* oil production during 1974–76. The pre-1974 period was excluded because modelling abandonment of state prorationing was not attempted, although that would be a useful and feasible extension. Several implications emerge. The controls significantly affected the supply elasticity of *U.S.* oil production. Prior to the *EPCA*, the elasticity of domestic production with respect to real world price was .09.[3] As hypothesized, the *EPCA* made that elasticity negative, −.48. Of course, these elasticities only measure instantaneous changes in production responding to increased real world price, and certainly understate final changes once all exploration and extraction responses are made.

TABLE 1—*Log* OF MONTHLY OIL PRODUCTION, 1,000 BARRELS/DAY (1974–76)

Explanatory Variable	Coefficient	*t*-statistic
Real World Oil Price	.472	2.58
EPCA Dummy Variable	.592	3.21
EPCA X World Oil Price	−3.034	−3.09
Entitlement Value	−.130	−0.87
Winter Dummy	.008	1.70
Spring Dummy	.004	0.74
Summer Dummy	−.000	−0.03
Time (months)	−.004	−8.84
Constant	9.051	252.06

Note: $R^2 = .9613$; $D.W. = 2.04$.

Changes in controls had only small impacts on domestic production levels. The entitlement program reduced production by under 1 percent, implying that ambiguous incentives under pre-*EPCA* controls were mostly offsetting for production by vertically integrated firms.[4] The *EPCA* did increase oil production but by only 1.5 percent.

The annual decay rate in domestic production during the 1974–76 period, estimated at 4.4 percent, was much higher than

[3]The elasticity of *U.S.* production with respect to a continuous variable equals the partial derivative of the equation with respect to that variable multiplied by the mean of that variable. For example, the supply elasticity prior to the *EPCA* equals .472 multiplied by $.1867 (the mean of the real world price per gallon variable).

[4]Entitlement program variable is zero prior to program, and equals the real value of the per barrel refining subsidy after the program (the average real subsidy was $.0423/gal.). The estimated impact of the program is the coefficient multiplied by that average subsidy. Data sources are available from the authors upon request.

experienced pre-1973, despite quadrupling of oil prices. Decay rates in the late 1960's and early 1970 fluctuated around zero, with some years actually showing growth. Final inferences must await future research that analyzes precontrol decay rates while accounting for abolition of prorationing.

IV. Conclusions

Government agencies have attempted to control only inframarginal production, while encouraging new production. Producers thereby received conflicting incentives to increase production, with the distinct possibility that initial oil production was higher because of controls. However, controls undoubtedly increased the decay rate in domestic production. After the 1975 *EPCA*, domestic production became *inversely* related to real world prices. Production was obtained at higher costs than without controls because the controls made marginal revenue different across different properties.

The final quantitative importance of these conclusions must await analysis of property-specific production data. Analysis of total *U.S.* production shows the impacts on production levels to be small initially, but impacts on the decay rate and supply elasticity were significant. Implications of these results, though preliminary, are thought provoking. The United States did not suffer significant reduced production from price controls during 1974–76. However, the controls have generated a perverse supply curve such that future increases in real world oil prices will transfer proportionately more *U.S.* income to *OPEC* than past price increases.

REFERENCES

R. E. Hall and R. S. Pindyck, "The Conflicting Goals of Energy Policy," *Publ. Int.*, Spring 1977.

Richard B. Mancke, *The Failure of U.S. Energy Policy*, New York 1974.

Charles E. Phelps and Rodney T. Smith, *Petroleum Regulation: The False Dilemma of Decontrol*, Santa Monica 1977.

DISCUSSION

STEPHEN BREYER, Harvard Law School: This interesting paper by Rodney Smith and Charles Phelps deals with "rent control" price regulation—a form of regulation used to control the price of housing and natural gas as well as oil. Such regulation seeks to transfer producer "rents" to consumers, typically by using a "tiered" pricing system, which sets a low price for the "old" low-cost product and a high price for the "new" high-cost product. Any such system must deal with two difficult sets of questions: Is the "high price" high enough to elicit adequate new production? How will the cheap "low price" product be allocated?

Smith and Phelps note several of the often conflicting complicated sets of incentives regulation can create as it struggles with these questions. They first point out that the "released barrel" program created an incentive to extract oil at a cost that exceeded the uncontrolled price. Second, integrated producers had an incentive to substitute controlled for uncontrolled (high price) oil in refining. Third, producers had an incentive to withhold oil in hope that controls would be relaxed; an incentive which became more powerful as the world price of oil increased. I shall comment briefly on each.

1. The discussion of the released barrel program may be less important empirically than is suggested. As Smith and Phelps note, whether it actually led producers to extract oil at a cost of, say, $12, $14 or $16 a barrel can better be determined by looking at individual properties, than by noting an overall production increase in 1974. The short life of the program, chaos and uncertainty in 1974, institutional delays in translating incentive into significant field investment, conservation laws or policies of companies that tie production rates to technical considerations, all would argue against a large output expansion due to released barrels.

The same discussion is more important than suggested in its implications for policy. The paper suggests the absurdity of using old oil rents in effect to finance the extraction of oil that costs more to produce than the uncontrolled price. Yet, the entitlements program, which took effect in 1974, embodies that same absurdity "writ large." By guaranteeing each refiner a proportionate share of cheap controlled oil, it not only makes controlled oil more expensive for those refiners who have it, but it makes imported oil cheaper for all the others. It offers refiners foreign oil, not at its incremential world price of, say, $14, but at an "average" (domestic and foreign) price of, say, $9 to $10. The result solves the "allocation" problem (it produces a market clearing price between the "low tier" price and the high foreign oil price). However it solves this problem by sacrificing the purpose of rent control price regulation, for in effect it uses old oil rents to finance the importation of high cost foreign oil. As Hall and Pindyck have pointed out, it transfers these rents not to consumers, but to the Arabs and, it encourages extra imports as well.

2. The discussion of the integrated refiner/producer's incentives rests upon the assumption that the price control system was totally ineffective; that refiners were able to capture the rents that price controls took from producers, and market prices for refined products were unaffected. This assumption, which Phelps and Smith have discussed elsewhere, strikes me as questionable, for the government imposed price controls on refined products that were based upon crude prices. How could these controls have been so ineffective as to allow refiners to capture rents that amounted to $6 or more per barrel? Their case seems to rest upon the facts that refined product is imported and sold at a profit, and that major refiners have "banked" costs since 1973 (which means that their resale price was lower than the price ceiling allowed them). The first of these facts has been disputed by others who state that the amount of foreign refined product imported (except for residual oil) is negligible. The second fact is insignificant

after Autumn 1974, when the entitlement program began. In late 1973 and early 1974, price controls would seem to have had some effect—at least if one can judge by the queues at gasoline stations. And the "banked cost" phenomenon during the summer of 1974 can be explained in part by lags, such as those stemming from retrospective price increases the Arabs imposed upon oil which had already been shipped and resold by the companies.

3. From a policy perspective the "withholding" phenomenon is troubling, for (as in the case of natural gas) it argues as much for definitively permanent controls as for decontrol. Measuring the effect of this incentive must also prove difficult, for it depends upon political calculation. Producers should have discounted heavily the possibility that they would ever receive an uncontrolled world price for existing low cost oil, for even proposals to decontrol tend to be coupled with proposals to tax "excess" profits. In fact, one could argue that the higher the world price, the less likely the public will allow the producers to retain this "excess" profit. This is not to deny the possible importance of politically related incentives. They play an important role in the natural gas debate and elsewhere. In the case of housing, for example, inadequate building, in the face of rent controls on old units but uncontrolled *new* unit rentals, can be in part explained by builder fear that rent controls will be extended to include units that once were new. Measuring the *importance* of the effect is the problem.

4. In general, I must confess, as a lawyer, to uncertainty about the extent to which the Smith and Phelps model demonstrates the empirical importance of the incentives they note. As they point out, they are forced to rely upon nationwide 1974 data. To what extent does the existence of radical price change, rapid regulatory change, institutional lags in adjustment, general turbulence in the industry weaken their empirical conclusion? I raise the question so that others, better qualified, can judge, but also, so that I can add that the value of their paper for those interested in regulation does not lie so much in its empirical conclusions as in its description of a series of incentives which were built into the price control program. This description provides additional grounds for believing that classical regulation cannot adequately deal with the two sets of questions listed at the beginning of this paper, and that other ways to deal with the problem should be sought.

A. MITCHELL POLINSKY, Harvard University: Most studies of the deterrent effect of capital punishment have concluded that there is none. Although this conclusion has been forcefully challenged by Isaac Ehrlich, a number of criticisms have been leveled against his contrary results. Given the unresolved nature of the debate, Kenneth Wolpin's new evidence in support of the deterrence hypothesis is of great interest. Using English time-series data, he is able to overcome two of the major problems raised with Ehrlich's time-series results based on *U.S.* data. In particular, Wolpin shows that his deterrence estimates are relatively robust with respect to a logarithmic or linear functional form and with respect to different time periods. Like Ehrlich, he also considers problems such as simultaneous equation bias and the confounding of incapacitative and deterrent effects. Overall, Wolpin's paper is carefully done and his results are presented judiciously.

Although Wolpin mentions several qualifications to his results in the longer version of his paper, his study will nonetheless be cited as establishing a deterrent effect. Moreover, his estimate of four victims saved for each additional execution of a convicted murderer will now be mentioned along with Ehrlich's corresponding numbers (which are nearly twice as large).

While Wolpin is able to overcome some problems, his estimates are still subject to a variety of other difficulties. Two in particular suggest that his results should be even more strongly qualified than he indicates. The first problem is that the lack of available data in average time served by

those convicted of murder and sentenced to imprisonment. If the average time served declines along with the probability of execution, then the deterrent effect of execution will be overstated. Since executions dropped substantially after 1956, it is especially important to know what happened after this year. On theoretical grounds, Wolpin argues (see his longer paper) that this omission could bias his results in either direction. On empirical grounds, he suggests that this problem may not be important. His only "evidence" is that the average length of court imposed sentences for aggravated assault, rape, and robbery increased between 1956 and 1968. The relevant question, of course, is not whether sentences increased, but whether actual time served increased. Given the rapid increase in all crimes after 1956, it is quite likely that jails became more congested because of the number of persons imprisoned. This would probably tend to reduce the average time served for all crimes. Although jail capacity could be expanded, it seems doubtful that it would have kept up with the number of persons imprisoned during a period of surprisingly rapid rises in crime rates.

The second problem concerns the omission of a variable representing the immigration of different racial groups into what would otherwise be a relatively racially homogeneous society. Wolpin notes (see his longer paper) that net immigration averaged 33,367 per year between 1956 and 1961, that it averaged only 15,307 per year between 1962 and 1966, and that the most rapid increases in the homocide rate occurred after 1961. From this he concludes that "[a]t least on the surface the homocide pattern does not conform to the immigration pattern." Wolpin's implicit "theory" seems to be that immigrants are more homocide prone for a short period after arriving, say, because of the absence of social stability. However, an alternative and equally plausible theory would lead to a quite different conclusion. Suppose that the homocide rate is partly a function of racial friction and that the level of racial friction is itself a function of the absolute number of, or the proportion of, the minority racial group (possibly with threshold effects). Such a theory would be quite consistent with the rapid rise in the homocide rate after 1961.

Each of these problems may account for the finding of a deterrent effect even if there is none. Since my alternative explanations are speculative, I do not mean to suggest that Wolpin's results are wrong. I only wish to stress his own observation that "there are important limitations of the data and difficult methodological issues yet to be resolved." His paper has, nonetheless, made a substantial contribution to our understanding of the deterrence hypothesis.

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS
OF THE
NINETIETH
ANNUAL
MEETING

NEW YORK, NEW YORK

DECEMBER 28–30, 1977

THE FRANCIS A. WALKER AWARD

Citation on the Occasion of the Presentation of the Medal to

SIMON KUZNETS

December 29, 1977

The American Economic Association awards, not more often than every five years, the Francis A. Walker medal to the economist "who has in the course of his life made a contribution of the highest distinction to economics." In 1977, this medal which represents the highest honor of the Association is given to Simon Kuznets.

Simon Kuznets: founder of modern national income and product measurement; designer of new systems of seasonal and cyclical measurement; discoverer of Kuznets cycles; frontiersman in economic demography; pioneer in quantitative studies of the economic growth of nations; explorer of income distribution. To few scholars is it given to make a fundamental difference to the development of his science. Kuznets has done so more than once. Economics, which is in his debt beyond its ability to acknowledge, hails him after fifty years of stunning achievement and prays only that its debt may continue to grow.

THE JOHN BATES CLARK AWARD

Citation on the Occasion of the Presentation of the Medal to

MARTIN S. FELDSTEIN

December 29, 1977

The John Bates Clark medal of the American Economic Association is awarded biennially "to that American economist under the age of forty who is adjudged to have made a significant contribution to economic thought and knowledge." Martin S. Feldstein has made many significant contributions, covering an astonishing array of economic methods and problems. His work spans a multidimensional spectrum embracing concrete policy issues, applied econometrics, statistical methods for econometrics, and economic theory. He pioneered in the economics of medical and hospital care. He has gone on to solve problems in benefit-cost analysis, public investment, taxation, social security, time preference and interest rates, asset holding under uncertainty, charitable donations, bequests, production functions, labor supply, unemployment, and inflation. The John Bates Clark medal is presented to Martin S. Feldstein for the high quality and extraordinary quantity and diversity of his achievements.

Minutes of the Annual Meeting
New York, New York
December 29, 1977

The Ninetieth Annual Meeting of the American Economic Association was called to order by President Lawrence Klein at 9:50 P.M., December 29, 1977, in the Grand Ballroom of the New York Hilton Hotel. He explained that he had written to the members a letter which they had not yet received. He then read the letter:

December 9, 1977

TO: Members of the American Economic Association

This has been a year of great losses in our professional ranks. It came as a great shock to learn of the sudden passing of our President-elect, Jacob Marschak, on July 26, 1977. Professor Marschak devoted unusual attention to his principal duty, the formulation of a program for the annual meeting of the Association, to be held in New York, December 27–30, 1977. Fortunately for the members of the Association, plans for an interesting and imaginative program were nearly complete by July. We are also fortunate that Jack Hirshleifer, a colleague of Jacob Marschak, worked closely with him in preparation of the program. Professor Hirshleifer has generously agreed to continue overseeing and shaping the program until the meetings are completed. Accordingly, the Association's Executive Committee has appointed Jack Hirshleifer as Honorary Program Chairman for the remainder of this year.

The Bylaws of the Association provide clearly for the succession to the office of presidency in case of incapacitation of the incumbent, but are not explicit on the rule of succession in case the President-elect becomes incapacitated. The Bylaws do state, however, that the Executive Committee has the power to fill vacancies in the list of officers. Under the advice of Counsel, that it was acting with proper authority, the Committee passed the following motions that:

Vice-President Robert Eisner be chosen Acting President-elect for 1977;

Tjalling Koopmans be chosen President for 1978.

Both Professor Eisner and Professor Koopmans have graciously agreed to serve the Association in their appointed duties for the remainder of this year and next year, respectively, in a sad and trying situation. Professor Eisner is experienced in affairs of the Association and will be able to carry out the remaining duties of President-elect. Professor Koopmans has been a close colleague, collaborator, and friend of Jacob Marschak for many years. He, more than anyone else, is in a position to carry out a term of presidency in the same spirit that Jacob Marschak would have done. A main task for the President is to prepare an address to be delivered to the annual meeting of the American Economic Association. Tjalling Koopmans is fully able, probably more so than anyone else, to deliver such an address in the spirit of Jacob Marschak's scientific career.

The Executive Committee appointed an *Ad-Hoc* Subcommittee, consisting of Burton Weisbrod and Robert Lampman, to review the provisions in the Bylaws relating to vacancies in the list of officers. They are to report to the December meeting of the Executive Committee.

Respectfully,
LAWRENCE R. KLEIN
President

The minutes of the meeting of September 17, 1976, were approved as published in the *American Economic Review Proceedings*, February 1977, pages 434–36.

The Secretary (C. Elton Hinshaw) presented the report of the Committee on Elections and the certification of the new

officers for 1978 as follows:

> In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee, the Electoral College, and the Committee on Elections.
>
> The Nominating Committee, consisting of Walter W. Heller, Chairperson, Nancy S. Barrett, Huey J. Battle, David M. Gordon, Bert G. Hickman, Irving B. Kravis, and Ralph W. Pfouts, submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-elect
Robert Solow

Vice-Presidents	*Executive Committee*
Edward F. Denison	Robert W. Clower
Joseph A. Pechman	William D. Nordhaus
William Vickrey	Lester C. Thurow
Phyllis A. Wallace	Marina v. N. Whitman

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. The Committee on Elections, consisting of Ben W. Bolch, Chairperson, Barbara Haskew, and C. Elton Hinshaw, *ex officio*, canvassed the ballots and filed results with the Secretary. From the report of the Committee on Elections, I have the following information:

Number of envelopes without names for identification	178
Number of envelopes received too late	86
Number of defective ballots	20
Number of legal ballots	5,122
	5,406

On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective offices:

President-elect (for a term of one year)
Robert Solow

Vice-Presidents (for a term of one year)
Edward F. Denison
Joseph A. Pechman

Members of the Executive Committee (for a term of three years)
Robert W. Clower
Marina v.N. Whitman

In accordance with the actions of the Executive Committee at its meeting on March 18, 1977, the amendment to Article I, Section 3 was submitted to the members in a mail ballot in conjunction with the balloting for officers. The ballots were canvassed by the Committee on Elections. On the basis of the canvass, I certify that the amendment was approved. As amended Article I, Section 3, now reads:

> Foreign economists of distinction may be elected honorary members of the Association. The Executive Committee is authorized to determine the number of foreigners to be elected honorary members. Past presidents of the Association and members who have been awarded the Walker Medal shall be Distinguished Fellows. Additional Distinguished Fellows may be elected, but no more than two in any one calendar year, from economists of high distinction in the United States and Canada. Candidates for Distinguished Fellowships shall be nominated by the Nominating Committee or the Executive Committee, and they shall be elected by the combined vote of the two committees. The Nominating Committee shall solicit and give due consideration to the recommendations of the Committee on Honors and Awards. The Nominating Committee is free to make no nominations in any particular year. However, it is not limited as to the number of candidates it may nominate in any year. Election to Distinguished Fellowship does not preclude election to any office of the Association.

The Secretary, Treasurer (Rendigs Fels), the Managing Editor of the *American Economic Review* (George H. Borts), the Managing Editor of the *Journal of Economic Literature* (Mark Perlman), and the

Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were available at registration and were also distributed at the meeting itself. (See their reports published in this issue.)

The Secretary presented the following resolutions, which were adopted unanimously:

> BE IT RESOLVED that this meeting record a special vote of thanks to the members of the 1977 Allied Social Science Associations' Local Arrangements Committee, chaired by Richard G. Davis, for their hard work, dedication, and efficient management of these meetings.

> WHEREAS Donald F. Turner has completed his term as Counsel of the Association, BE IT RESOLVED that this meeting express its warmest appreciation for the outstanding contribution he has made to the Association through his perceptive and judicious advice and counsel.

> BE IT RESOLVED that this meeting record a special vote of appreciation to Jack Hirshleifer for his role in planning a program intellectually stimulating, appropriately varied, and high in quality.

> At these meetings we mourn the loss of our esteemed colleague and President-elect, Jacob Marschak, and, at the same time, recognize the magnificent work that he accomplished in putting the 1977 program together. It has been an outstanding program that clearly reveals the unusual and perceptive insights of Jacob Marschak into new and broader directions for our subject. The interdisciplinary and frontier areas of economics were fully evident. Therefore, BE IT RESOLVED that this meeting record a vote of recognition for Jacob Marschak who chaired the organization of this year's meeting.

There being no old business, the President called for new business. He stated that five resolutions by members had been submitted to the Secretary a month in advance of the meeting as required by the Association's regulations. These resolutions had been distributed before and at the meeting.

The President called for discussion of the resolution submitted by Eric D. Bovet and Ralph Z. Politte. Since no one spoke in behalf of the resolution, it was moved, seconded, and PASSED that the resolution be tabled indefinitely.

The President called for discussion of the resolution submitted by Betty A. Little and Ella F. Filippone. Since no one spoke in behalf of the resolution, it was moved, seconded, and PASSED that the resolution be tabled indefinitely.

The President called for discussion of the resolution submitted by Huey J. Battle and Marcus Alexis. The resolution read:

> WHEREAS the supply of minority group members in the economics profession continues to be depressingly low in relation to the minority population, and whereas the Committee on the Status of Minority Group Members in the Economics Profession has been successful in securing funds and offering a predoctoral summer program for minority students which has greatly expanded opportunities for said students to prepare themselves for graduate work in economics, and whereas funding is increasingly difficult to obtain;

> BE IT RESOLVED that the American Economic Association reaffirm its commitment to increasing the supply of minority group members in the economics profession and that the American Economic Association will continue to seek funding for the Summer Program in Economics for Minority Students and to commit its own resources where necessary until the flow of minority group students through graduate programs in economics approaches the relative share(s) of minority group(s) students.

Klein expressed the concern of the Executive Committee that the resolution represented a "blank check" for the summer program for minority students. Marcus

Alexis spoke in favor of the resolution. He stated that the intent was not to give a blank check to the program but to assure the continuation of the program by guaranteeing back-up financing in the event that outside funds could not be obtained. Each of the four previous summer programs had been operating on a year to year financial basis. Last summer the students and the money arrived simultaneously. Long-term (at least three years) funding is desirable for planning and management purposes.

Bernard Anderson asked if the phrase, "the American Economic Association will continue to seek funding. . . ," meant that the officers of the Association would actively engage in helping raise funds for the program. Alexis responded that that was the intent and that the officers had been very helpful in the past.

Fels moved that the resolution be amended by striking the words "where necessary until the flow of minority group students through graduate programs in economics approaches the relative share(s) of minority group(s) students" and adding the words "for that purpose." Robert Wolfson suggested the insertion of the word "from" between "commit" and "its own resources." Fels accepted the change, and there was no objection from the floor.

Alexis moved, and it was seconded, that the resolution be amended further by adding a third paragraph:

> BE IT FURTHER RESOLVED that the American Economic Association underwrite the necessary matching funds for the 1978 summer program for minority students in the event that external funding is not found.

Alexis's motion to amend failed.

Klein then called for the vote on the resolution as amended by Fels. The resolution as amended PASSED. It reads:

> WHEREAS the supply of minority group members in the economics profession continues to be depressingly low in relation to the minority population, and whereas the Committee on the Status of Minority Group Members in the Economics Profession has been successful in securing funds and offering a predoctoral summer program for minority students which has greatly expanded opportunities for said students to prepare themselves for graduate work in economics, and whereas funding is increasingly difficult to obtain;
>
> BE IT RESOLVED that the American Economic Association reaffirm its commitment to increasing the supply of minority group members in the economics profession and that the American Economic Association will continue to seek funding for the Summer Program in Economics for Minority Students and to commit from its own resources for that purpose.

The President called for a discussion of the resolution submitted by Alfred Kraessel and George Tzannetakis. Since no one spoke in behalf of the resolution, it was moved, seconded, and PASSED that the resolution be tabled indefinitely.

Before calling for discussion of the resolution submitted by Robert Cherry and Patrick Clawson, the President stated that Counsel had advised that the resolution is in conflict with Paragraph Third, Subparagraph 3, of the Association's certificate of incorporation,[1] and is accordingly out of order. The resolution reads:

> WHEREAS, sociobiology is a political ideology associated with a long history of legitimating reactionary policies;
>
> WHEREAS, the American Economic Association, through the *Journal of Economic Literature*, the Annual Meetings, and forthcoming *Papers and Proceedings*, has embarked on a one-sided program of legitimating sociobiology as a science;
>
> WHEREAS, there are a number of economists who have presented papers and published articles on the antisocial, antiscientific nature of sociobiology, but were not invited to par-

[1]For the Certificate of Incorporation, see page ix of the 1974 *Directory of Members*.

ticipate in the 1977 program;

BE IT RESOLVED: The American Economic Association shall

1. Cancel its plans to publish the papers from the "Economics and Biology" session in the *Papers and Proceedings*.
2. Select a chairperson for a session at the August 1978 meetings to guarantee that the antisociobiology position will be articulated.
3. Publish the papers from the session at the August 1978 meetings in the *Papers and Proceedings*.
4. Encourage the submission of articles to the *Journal of Economic Literature* which would identify the role of sociobiology within and without the economics profession. It would guarantee that the best of these articles submitted by January 1979 will be published.

Cherry spoke in favor of the resolution and against the Association's tradition of allowing the Program Chairman (the President-elect) to select which sessions will be published in the *Papers and Proceedings*. He argued that the session in question, "Economics and Biology," was one-sided and that procedures should be developed to insure balanced presentations in sessions dealing with controversial issues.

At this point, the Chair ruled the resolution out of order. Cherry appealed the Chair's ruling. Clawson spoke in favor of the resolution and supported Cherry's appeal. He stated that he considered it no accident that the session was one-sided; that it reflected a deliberate attempt to exclude a particular political view. Leo Raskind, new counsel of the Association, advised that Item 1 of the resolution may be inconsistent with one of the Association's expressed objectives, "The encouragement of perfect freedom of economic discussion." (See Paragraph Third, Subparagraph 3 of the Certificate of Incorporation.) James Kinney disagreed and advocated overruling the Chair's decision. Robert Eisner spoke against the resolution, but wished it had not been ruled out of order. Abba Lerner stated that adopting Item 1 of the resolution would constitute a form of censorship. A motion to sustain the Chair's ruling that the resolution was out of order was made, seconded, and PASSED.

At this point President Klein introduced the new president, Tjalling Koopmans, who took the chair. There being no further business, the meeting was adjourned at 11:55 p.m.

C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in Arlington, Virginia, March 18, 1977.

The first meeting of the 1977 Executive Committee of the American Economic Association was called to order at 9:20 A.M. on March 18, 1977, in the Arlington Hyatt House, Arlington, Virginia. Present as members of the Executive Committee were Lawrence R. Klein, presiding, Carolyn Shaw Bell, George H. Borts, Robert Eisner, Rendigs Fels, C. Elton Hinshaw, Anne O. Krueger, Robert J. Lampman, Jacob Marschak, Franco Modigliani, Marc Nerlove, Mark Perlman, Edmund S. Phelps, Alice M. Rivlin, and Burton A. Weisbrod. Donald F. Turner, the Association's counsel, was present for part of the meeting. Present for part of the meeting as members of the Nominating Committee were Walter W. Heller, chair, Huey Battle, Nancy S. Barrett, David M. Gordon, Bert G. Hickman, Irving B. Kravis, and Ralph W. Pfouts. Present for part of the meeting as members of the Committee on Honors and Awards were Irma Adelman, chair, Moses Abramovitz, Marcus Alexis, John S. Chipman, and Carl Christ. Present for part of the meeting as guests were Barbara Reagan and Wilma St. John.

Minutes. The minutes of the meeting of September 15, 1976 were approved.

Report of the Secretary (Hinshaw). The Secretary reported that the total number of registrants at the 1976 Atlantic City meetings was 4,179. There were 282 scheduled events: 189 sessions, 76 fee events and meal functions, and 17 miscellaneous functions. The 1977 annual meetings will be held in New York on December 28–30, with the employment service beginning operations on the 27th. The schedule for subsequent meetings is: August 29–31, 1978 in Chicago; December 28–30, 1979 in Atlanta; September 5–7, 1980 in Denver; and December 28–30, 1981 in Washington, D.C.

Because the 1976 annual meeting came at an early date in the academic year, a second, supplementary placement meeting was held January 7–9, 1977 at the O'Hare Hilton in Chicago, Illinois. Almost 1,400 people registered (1,036 applicants and 361 employers) and approximately 1,300 attended, roughly the same number as in Atlantic City. The unexpected size of the market created many problems, but the market was able to overcome the Secretary's forecasting and planning. After the initial shock, employers and applicants adjusted to the constraints of too few staff, too few binders, and an inadequate communications system.

Both the Chairpersons Group and the Placement Officers Group have passed resolutions calling for only one placement meeting each year. In those years when the annual meeting occurs in December, the labor market should be scheduled to coincide with it. In those years when the annual meeting is scheduled around the Labor Day period, the labor market should be scheduled in late December or January. For the annual meeting scheduled for August 29–31, 1978 in Chicago, the Executive Committee VOTED to provide no formal placement service at that meeting but to organize the official employment service for late December or early January. The Secretary is to announce the specific dates and site as soon as arrangements can be made.

The Secretary reported that twenty-five life members have contributed a total of $1,602. Their names are:

Mariano Alierta	Andrea Maneschi
William J. Baumol	W. W. McPherson
M. Gardner Clark	Frederic L. Pryor
Edna Douglas	James S. Raji
John Eddison	Jung Han Rhi
Edwin Flynn	Lawrence Ritter
David C. Gogerty	Richard Scheuch
Walter W. Heller	Jeronimo Sotillo
Norris O. Johnson	William W. Stevenson
Kenzo Kobayashi	Ralph L. Thomas
Frank J. Kottke	William Vickrey
Orlando H. Lobo	Henry H. Villard
Spiro J. Lotsis	

It was VOTED to include a general request for contributions in the dues invoices sent to all members and subscribers.

Report of the Treasurer (Fels). The financial position of the Association is the best it has been since the end of 1969. The surplus for 1976 was $150 thousand with another surplus projected for 1977. Rising costs, however, can be expected to wipe out the surplus in the next few years even if no new programs of expenditure are undertaken. In view of the relentless increase in costs, the present favorable financial position will be temporary unless expenditures are restrained or dues increased. It was VOTED to approve the 1977 budget submitted by the Treasurer and the Budget Committee. (For details of the budget, see the Treasurer's Report in the September 1977 issue of the *American Economic Review*.) It was VOTED to increase dues and subscriptions up to 5 percent effective January 1, 1978.

Report of the Editor of the American Economic Review (Borts). On recommendation of the Managing Editor, the following persons were elected to the Board of Editors of the *American Economic Review* for three-year terms: Albert Ando, Elizabeth Bailey, Frederic M. Scherer, and David Bradford. In addition, Martin Feldstein and Jerome Stein were reelected to second terms. The Editor reported that, beginning in 1978, the editorship of the *Papers and Proceedings* issue of the *AER* is being transferred from the Secretary-Treasurer's office to his office. It was VOTED to allow the addition of 48 pages to the space devoted to the "*Papers*" section of the *Papers and Proceedings* for 1978.

Report of the Editor of the Journal of Economic Literature (Perlman). The Editor reported that the 1972 and 1973 volumes of the *Index of Economic Articles* are scheduled for publication this year and the 1974 and 1975 volumes are expected in 1978. Thereafter, one volume each year will be published. He expressed dissatisfaction with the current arrangement for distribution of the *Index* and encouraged the Association to consider other distributors.

Report of the Director of Job Openings for Economists (Hinshaw). The Director reported on the number of employers listing jobs during 1976, the number of openings, and the fields of specialization most in demand. (For details of the report, see the April 1977 issue of *Job Openings for Economists*.) Unaudited figures for revenues and expenses indicate a deficit of $3,200 in 1976. The Association's auditors have advised that *JOE* is an "unrelated business" and subject to federal income tax. In the past no overhead was allocated. In 1976 and in the future, overhead will be charged to *JOE*. Close scrutiny of costs should be sufficient to maintain an accounting deficit in the coming years and allow *JOE* to operate without a budgeted subsidy from the Association. The Director was urged to be more aggressive in seeking nonacademic listings.

Committee on Honors and Awards (Adelman). The Electoral College consisting of the Executive Committee and the Committee on Honors and Awards meeting together VOTED to award the Francis A. Walker Medal to Simon Kuznets and the John Bates Clark Medal to Martin Feldstein. It was VOTED that to be eligible for the John Bates Clark Medal, one must be under 40 during the entire year in which the award is made.

Nominating Committee (Heller). The Electoral College consisting of the Nominating Committee and the Executive Committee meeting together chose Robert Solow as the nominee for President-elect and Leonid Hurwicz and Harry G. Johnson as Distinguished Fellows. The chairperson of the Nominating Committee reported the following nominees for other offices in the 1977 election: for Vice President (two to be chosen), Edward F. Denison, Joseph Pechman, William Vickrey, and Phyllis Wallace; members of the Executive Committee (two to be chosen), Robert W. Clower, William D. Nordhaus, Lester Thurow, and Marina v. N. Whitman.

Committee on the Status of Women in the Economics Profession (Reagan). The chairperson reported that the Committee wished to charge dues to those wishing to

become associate members of the group. In the past, the Committee had requested donations from those wishing associate membership status. The Executive Committee concurred with the proposal to establish dues. Reagan then reviewed the history of the publication of papers given at the annual meetings under the Committee's aegis. She stated that the sessions had been published in 1973, 1974, and 1975. Only one paper was published in 1976, and, currently, none of the 1977 papers are to be published. She suggested that the Executive Committee establish a publication policy for the Committee's session akin to that for the Committee on Economic Education: a five-year right to be published provided the chairperson of the Committee judged the papers to merit publication.

National Economic Association (Battle). Battle reviewed the publication history of joint NEA and AEA sessions at the annual meetings. He stated that the 1977 program does not currently provide for publication of NEA-AEA joint sessions. He recommended that the Executive Committee adopt a policy of publishing all papers presented at the joint session.

Having previously approved an additional 48 pages for the "Papers" section of the *Papers and Proceedings*, the Executive Committee voted to postpone discussion of a publication policy for papers presented at the annual meetings. It was understood that the Program Chairperson would continue to select the papers to be published.

Committee on the Status of Minority Groups in the Economics Profession (Alexis). The chairperson reported that he is currently seeking long-term financing for the summer training program. He is hopeful that a three-year grant will be awarded.

Committee on Political Discrimination. Klein reported the resignation of F. Ray Marshall as chairperson and announced the appointment of Gerald Somers as his replacement. Gordon expressed disappointment with the lack of progress made by the Committee during the last year and a half and indicated that those most concerned will examine their alternatives. It was VOTED that the President request the Committee to complete the transfer of the case by case investigative function to the American Association of University Professors as soon as possible, to determine the feasibility of conducting a research project on political discrimination and report its decision to the Executive Committee by May 1, 1977, and, if a research project is feasible, to develop a proposal for conducting such research by November 1, 1977.

Term of Counsel (Klein). The President reported that Turner's term as counsel to the Association expires December 31, 1977 and that he does not wish to be reappointed. The President will report on a replacement at the December meeting. The Executive Committee expressed its appreciation and gratitude to Turner for his outstanding service to the Association.

1977 Program (Marschak). The Program Chairman reported that plans for the invited paper sessions were complete and that 132 abstracts of contributed papers had been received. The contributed papers will be categorized by topic and organized into sessions.

Committee on U.S.–Soviet Exchange. The written report of the Chairman, Lloyd G. Reynolds, stated that the Committee had drafted a request to the National Science Foundation for a two-year grant to cover a visit of Soviet economists to the United States in 1977 and a return visit of American economists to the Soviet Union in 1978. The tentative date of the 1977 Soviet visit is October 9–23. The report also stated that the National Bureau of Economic Research has been conducting a substantial exchange program funded by NSF. The cooperating institutions on the Soviet side are the Central Economic Mathematics Institute and the Research Institute of Gosplan. The National Bureau is interested in transferring some of these activities to AEA auspices. The President appointed Eisner to communicate with CEMI and investigate what role the AEA might play in the exchanges. It was VOTED to empower the President to investigate the possibility of the Association accepting the responsibility for exchanges

and make a sensible arrangement concerning exchange procedures.

Survey of Financial Characteristics of Consumers (Bell). Bell reported that Arthur F. Burns had rejected her suggestion that the Federal Reserve System initiate another study of the financial characteristics of consumers. It was VOTED that the President should explore with the Board of Governors the possibility of establishing an AEA advisory committee to the Federal Reserve System on research and data collection.

Federal Funding of Basic Research (Modigliani). Modigliani reported that the Committee recommends that the funding process be monitored by a member of the Executive Committee who would act as liaison between the Association and federal agencies and make recommendations as seemed appropriate. Nerlove, the Executive Committee member appointed to the task, stated that it is impossible for one person to monitor the entire research funding activities of all government agencies. If the National Science Foundation would monitor the funding of other agencies, he would be liaison between NSF and the Association. The President is to explore the possibility of NSF cooperating with the Association in such monitoring.

Economics Institute (Krueger). In the absence of Wyn Owen, Krueger presented the Institute's request for an appropriation of $8,500 for fellowship funds. It was VOTED to lend $8,500 to the Institute with the expectation that when other funds are received the loan will be repaid.

Finance Committee. The written report of the Chairman, Beryl W. Sprinkel, stated that the Committee had decided to maintain the equity proportion of the portfolio at approximately two-thirds while permitting the flexibility to vary the ratio 5 percentage points in either direction.

Search Committee for Editors (Klein). The President asked for advice concerning the composition and charge of the Committee.

Bylaw Revision (Hinshaw). An anomaly remains in the bylaws. In Article I, Section 3, the third sentence says that the Executive Committee may elect Distinguished Fellows, but the next sentence specifies election by a combined vote of the Nominating and Executive Committees. It was VOTED to submit to the members for approval by mail ballot the following amendment to the bylaws: Change Article I, Section 3, to read:

> Additional Distinguished Fellows may be elected, but not more than two in any one calendar year, from economists of high distinction in the United States and Canada. Candidates for Distinguished Fellows shall be nominated by the Nominating Committee or the Executive Committee, and they shall be elected by the combined vote of the two committees.

Terms of Secretary and Treasurer (Klein). The current terms of the Secretary and the Treasurer expire December 31, 1978. It was VOTED to reappoint both for another three-year term.

Date of Spring Meeting. It was agreed that the Executive Committee would meet on March 17–18, 1978, possibly in Nashville, Tennessee.

The meeting adjourned at 10:40 p.m.

Minutes of the Meeting of the Voting Members of the Executive Committee in Washington, D.C., November 6, 1977.

The special meeting of the voting members of the 1977 Executive Committee of the American Economic Association was called to order at 1:10 P.M. on November 6, 1977 in the Stouffer's Center Hotel, Washington, D.C. Present as voting members of the Executive Committee were Lawrence Klein, presiding, Carolyn Shaw Bell, R. A. Gordon, Robert Lampman, Franco Modigliani, Marc Nerlove, Edmund Phelps, Alice Rivlin, and Burton Weisbrod. Absent were Robert Eisner and Anne Krueger. Present as guests were C. Elton Hinshaw, Secretary, and Donald Turner, Counsel.

Klein announced that he had called the meeting to consider what actions the Executive Committee should take as a result of the death of Jacob Marschak, 1977

President-elect of the Association. Section IV, Article 5, of the bylaws provides that the Executive Committee "may fill vacancies in the list of officers" (page xi, *1974 Directory of Members*). After seeking advice from counsel about interpretations of the bylaws, the Executive Committee VOTED to elect Tjalling Koopmans President of the Association for 1978, Robert Eisner Acting President-elect for 1977, and Jack Hirshleifer Honorary Program Chairman for 1977. It was decided that Klein should write a letter to the members explaining the Committee's actions.

Klein appointed Weisbrod and Lampman to an *Ad Hoc* Subcommittee of the Executive Committee to consider amending the bylaws relating to vacancies in the list of officers. The Subcommittee is to report to the December 27, 1977 meeting of the Executive Committee.

It was agreed that the Executive Committee would meet on March 17 and 18, 1978 in Washington, D.C. or New York.

The meeting adjourned at 2:35 p.m.

Minutes of the Meeting of the Executive Committee in New York, New York, December 27, 1977.

The third meeting of the 1977 Executive Committee was called to order at 10:10 A.M. on December 27, 1977 in the New York Hilton Hotel, New York, New York. The following members were present: Lawrence R. Klein presiding, Carolyn Shaw Bell, George H. Borts, Robert Eisner, Rendigs Fels, Robert Aaron Gordon, C. Elton Hinshaw, Anne O. Krueger, Robert J. Lampman, Franco Modigliani, Marc Nerlove, Mark Perlman, Edmund S. Phelps, and Burton A. Weisbrod. Also present were newly elected members Robert W. Clower, Edward F. Denison, Tjalling Koopmans, Joseph A. Pechman, Robert M. Solow, and Marina v. N. Whitman. Present as counsel were Donald F. Turner and Leo J. Raskind. Present as guests for parts of the meeting were Marcus Alexis, Ann Friedlaender, Gary Fromm, David M. Gordon, Martin Landsberg, Lloyd G. Reynolds, Wilma St. John, and Gerald Somers.

Klein introduced Raskind as the new counsel of the Association and thanked Turner, the outgoing counsel, for his sagacious and wise advice during his years of service. The Executive Committee VOTED a special note of gratitude to Turner and expressed its appreciation to him with a round of applause.

Minutes. The minutes of the meetings of March 18, 1977 and November 6, 1977 were approved.

Report of the Secretary (Hinshaw). The Secretary reported that the 1978 annual meetings will be held in Chicago on August 29–31 and a separate placement service is scheduled for December 28–30 in Chicago with the Conrad Hilton Hotel serving as headquarters. The Executive Committee VOTED to accept the Secretary's recommendation to continue to regard the placement service as part of the activities of the annual meetings and to finance it from the revenues of those meetings. The Executive Committee VOTED to approve the Secretary's recommendation of New York City as the site of the 1982 meetings and instructed the Secretary to delay as long as feasible the decision between the Labor Day period or the Christmas-New York period as dates for the meeting. The Secretary announced plans to publish a handbook in 1978 and sought advice regarding a change in the material contained in the handbook and the handbook questionnaire. It was VOTED to appoint a committee to design the questionnaire and to empower it to act on behalf of the Executive Committee.

Report of the Treasurer (Fels). The Treasurer presented preliminary financial reports for 1977. (For details of the reports, see the Treasurer's Report in these *Proceedings*.) He stated that the financial position of the Association had improved dramatically in 1976 and 1977 and projected a substantial surplus of $181,000 in 1978. The Executive Committee approved the budget submitted by the Treasurer and the Budget Committee without a formal vote.

Report of the Editor of the American Economic Review (Borts). The Editor reported that twenty-five sessions are scheduled for publication in the forthcom-

ing *Papers and Proceedings*. He stated that revenues from page charges were lower than expected. It was suggested that authors be informed that the page charge is expected to be paid if a grant or foundation is acknowledged as funding the research. It was VOTED to publish the *Papers and Proceedings* in May of each year regardless of the dates of the annual meeting.

Report of the Editor of the Journal of Economic Literature (Perlman). On recommendation of the Editor, the following persons were elected to the Board of Editors of the *Journal of Economic Literature:* David Laidler, Harvey Leibenstein, Daniel McFadden, and Roger Ransom. He reported that *Annual Indexes* for the years 1972 and 1973 were published in 1977. The next three volumes will appear during the next two years.

Report of the Director of Job Openings for Economists (Hinshaw). The Director reported that, in an effort to increase the number of nonacademic job listings, he had written to over 200 corporations and government agencies inviting them to list their vacancies. Academic jobs continue to constitute about two-thirds of all vacancies advertised.

Committee on the Status of Minority Groups in the Economic Profession (Alexis). Alexis reported that the Sloan Foundation had awarded a grant to the Association to support the summer program for minority students who are interested in pursuing the Ph.D. in economics. The grant requires matching funds. He also reported receiving a planning grant from the Ford Foundation to establish a consortium of five universities to attempt to increase the number of minority students undertaking graduate training in economics. The Executive Committee VOTED a special note of appreciation to Alexis for his effective management of the program.

Committee on U.S.-Soviet Exchanges (Reynolds). Reynolds reported on the November 27-December 11, 1977 visit of six Soviet economists to the United States for a conference on "Aspects of Labor Economics." It was VOTED to continue the exchange for four more years and attempt to promote research cooperation of a more serious nature than can be accomplished in brief symposia.

Committee on Political Discrimination (Somers). Somers reported that the Committee recommended the following procedure for handling individual cases: (1) Members of the AEA should continue to file complaints with the Committee, but these complaints should be referred to the American Association of University Professors (AAUP). (2) The Committee should work closely with the AAUP to help expedite investigations of such complaints. (3) The Committee should review decisions made by the AAUP and advise members of other available channels. (4) The Committee should not contact university officials on anything other than the most routine details while the case is under consideration by the AAUP.

The Committee also recommended that the Association sponsor and fund a research project on patterns of political discrimination. It was VOTED to allocate $10,000 to the Committee to be used at its discretion for a study on political discrimination. It was understood that the Committee on Political Discrimination would seek additional funding from other sources. It was suggested that the Committee discuss with the Secretary the possibility of coordinating a survey questionnaire with the handbook questionnaire.

Future Convention Sites (Friedlaender). On behalf of the Committee on the Status of Women in the Economics Profession and other concerned members of the Association, Friedlaender proposed that the Association change the sites for the 1978 and 1979 meetings because they are currently scheduled for states (Illinois and Georgia) which have not ratified the Equal Rights Amendment to the U.S. Constitution. Counsels Turner and Raskind advised that such an action would conflict with Article 3 of the Association's Certificate of Incorporation—"The Association as such will take no partisan attitude, . . ." After a

consideration of the possible economic and legal consequences of moving the sites, it was moved that immediate inquiries be made to determine the feasibility and desirability of withdrawing from Atlanta in 1979. The motion failed. It was understood that this action did not imply any position on the Equal Rights Amendment.

Bylaws (Weisbrod). The *Ad Hoc* Committee to Review Procedures for Replacement of President and President-elect, consisting of Lampman and Weisbrod, reported on various procedures that might be adopted to fill vacancies in the offices of President and President-elect. It was VOTED that the *Ad Hoc* Committee should recommend a specific proposal to the next meeting of the Executive Committee.

1978 Program (Solow). The Chairman gave a brief report on plans for the 1978 program.

Joint Ad Hoc Committee on Government Statistics (Denison). Denison reported that the Joint *Ad Hoc* Committee, consisting of representatives from nine professional associations, recommended: (1) There be established a Committee of Professional Associations on Federal Statistics. (2) There be established an office of Professional Associations on Federal Statistics under the policy guidance of the Committee. (3) Each participating Association contribute $2,000 a year to the costs of operation of the office for an initial three-year trial period. It was VOTED to join the Committee of Professional Associations on Federal Statistics and appoint a representative to the Committee.

Federal Funding of Economic Research (Fromm). Following the recommendations of Fromm, the Executive Committee VOTED to approve the formation of a new *ad hoc* committee to report on federal support for economic research during the past decade, activities undertaken by other professional societies in support of research funding, and actions that the AEA should consider germane to the encouragement of economic education and research.

Economists in Argentina (Klein). The President introduced a resolution submitted by Wassily Leontief. It read:

WHEREAS:

The members of the Executive Committee of the American Economic Association have been informed that:

(a) The Government of Argentina has imprisoned and has continued to detain five of our Argentine colleagues from the National University of the South, accusing them of "ideological and social-cultural infiltration."

(b) The principal specific charge against the accused is that they participated in creating a program leading to a degree in economics, said program being similar in content to that in universities in the United States and other industrialized democracies.

(c) The alleged crime is also based on the charge that the accused encouraged their students to continue their studies in different universities around the world, including the University of Colorado, which offers under the sponsorship of the American Economic Association summer courses to students from abroad.

THEREFORE, BE IT RESOLVED THAT:

1. The Executive Committee of the American Economic Association condemns this reported attack on the academic activities of its Argentine colleagues; and urges (in the strongest possible terms) that any such charges against them be dismissed.

2. The President is directed to communicate this resolution to the Government of Argentina in such manner and through such channels as he may determine to be appropriate.

It was VOTED to direct the President to write a letter to the Government of Argentina expressing concern and seeking additional information about the situation.

Resolutions from Members (Klein). The Executive Committee reviewed and discussed the resolutions submitted by members for consideration at the annual business meeting.

Committee on Editorships (Klein). In the

absence of James Tobin, Chair of the Committee, Klein reported. Acting on the committee's recommendation, the Executive Committee VOTED to offer Naomi Perlman a four-year term as Associate Editor of the *Journal of Economic Literature* to manage the indexing and bibliographic functions. The term ends December 31, 1981, but can be renewed. It was understood that the new Managing Editor will have the same authority as the current one and can make, subject to his responsibilities to the Executive Committee, policy decisions about the whole journal, including indexing and bibliographic functions.

Finance Committee (Eisner). Eisner reported that the committee had voted to instruct Stein, Roe & Farnham, investment counsel, to observe limits of 50 and 67 percent equities for the Association's total portfolio.

Date of Spring Meeting. It was agreed that the Executive Committee would meet on March 17–18, 1978 in New York City.

The meeting was adjourned at 11:45 p.m.

C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1977

Annual Meetings. In 1978 the annual meetings will be held at the Conrad Hilton Hotel in Chicago, Illinois on August 29–31. The schedule for subsequent meetings is: December 28–30, 1979, in Atlanta with headquarters at the Atlanta Hilton Hotel; September 5–7, 1980, in Denver with headquarters at the Denver Hilton Hotel; and December 28–30, 1981, in Washington, D.C. with headquarters at the Washington Hilton Hotel. The Executive Committee has selected New York as the site of the 1982 meetings but has not decided about the dates. It has also made a tentative decision to meet in San Francisco, December 28–30, 1983.

Employment Services. Because the 1978 annual meetings occur at an early date in the academic year, it was decided to provide the employment services at a later time. The 1978 employment center is scheduled for December 28–30 in Chicago with the Conrad Hilton Hotel serving as headquarters. Formal employment services will not be provided at the annual meetings in August.

The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service under the direction of Kathy Nichols. All economists looking for jobs and employers are urged to register. This is a placement service which maintains the anonymity of employers. The Association is indebted to Ms. Nichols not only for the Registry but also for her and her staff's assistance and supervision of the employment service provided at the annual meetings.

Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and of their professional obligation to list their openings.

Membership. The total number of members and subscribers, shown in Table 1, reached an all-time high of 26,787 at the end of 1975. The introduction of a progressive dues structure in 1976 may account for most of the decline in the numbers of members and subscribers in 1976 and 1977. The 1977 count is the first provided by the new computerized system. It is probable that "cleaning" the files during the transition accounted for a small part of the decrease.

Permission to Reprint and Translate. Official permissions to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 236 in 1977 compared to 200 in 1976. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

Visiting Economics Scholars Program. The Visiting Economics Scholars Program has continued under the direction of the Secretary. Its purpose is to facilitate visits by leading economists to smaller colleges emphasizing teaching. The colleges are ex-

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1975	1976	1977
Class of Membership			
Annual	16,011	15,102	14,379
Junior	2,367	2,631	1,731
Life	383	399	375
Honorary	19	36	33
Family	335	344	284
Complementary	449	560	537
Total Members	19,564	19,072	17,386
Subscribers	7,223	7,134	**6,728**
Total Members and Subscribers	**26,787**	**26,206**	**24,114**

pected to pay part or all of the costs of the visits; at a minimum they take care of the local expenses and travel costs of the visitor. During 1976–77 there were two such visits.

Computerization of the Membership–Subscribers File. The membership–subscriber file was converted from manual records to a computer system during 1977. The master file is now maintained by Epsilon Data Management, Inc. in Boston, Massachusetts.

Committees and Representatives. Listed below are those who served the Association during 1977 as members of committees or as representatives. The year in parenthesis indicates the final year of the term to which they have been appointed most recently. On behalf of the Association, I wish to thank them all for their services.

Standing Committees

ADVISORY COMMITTEE ON THE HISTORY OF THE ASSOCIATION
George J. Stigler, *Chair*
Joseph Dorfman
Harold F. Williamson, *Corresponding Secretary*

ADVISORY COMMITTEE ON STUDIES OF THE LABOR MARKET FOR ECONOMISTS
F. Ray Marshall, *Chair*
Barbara Reagan
T. Aldrich Finegan
Francis M. Boddy

BUDGET COMMITTEE
Burton A. Weisbrod, *Chair* (1977)
Edmund S. Phelps (1978)
Robert J. Lampman (1979)
Lawrence Klein, *ex officio* (1977)
Rendigs Fels, *ex officio*

CENSUS ADVISORY COMMITTEE
James R. Nelson, *Chair* (1979)
Burton Malkiel (1979)
Arnold Zellner (1979)
Andrew F. Brimmer (1979)
Norman Simler (1977)
George L. Perry (1977)
Carolyn Shaw Bell (1979)
Lee Preston (1977)
Phyllis Wallace (1977)
Dale Jorgenson (1977)
Barbara Bergmann (1978)
Robert F. Lanzillotti (1978)
William Niskanen (1978)
Anne P. Carter (1978)
Richard Ruggles (1978)

COMMITTEE ON ECONOMIC EDUCATION
Allen C. Kelley, *Chair* (1979)
Walter Heller (1979)
Phillip Saunders (1977)
Elizabeth Allison (1978)
John Siegfried (1978)
W. Lee Hansen (1978)
Rendigs Fels, *ex officio*

COMMITTEE ON HONORARY MEMBERS
Leonid Hurwicz, *Chair* (1980)
William Baumol (1982)
Hollis B. Chenery (1982)
Bent Hansen (1978)
W. Arthur Lewis (1978)
Paul A. Samuelson (1980)

COMMITTEE ON HONORS AND AWARDS
Irma Adelman, *Chair* (1978)
Carl F. Christ (1982)
Marcus Alexis (1978)
John Chipman (1980)
James W. McKie (1980)
Moses Abramovitz (1982)

COMMITTEE ON THE STATUS OF MINORITY GROUPS IN THE ECONOMICS PROFESSION
Marcus Alexis, *Chair*
George Borts
Andrew Brimmer
Alice Rivlin
James Tobin
Charles Z. Wilson

COMMITTEE ON POLITICAL DISCRIMINATION
Gerald G. Somers, *Chair* (1978)
Kenneth J. Arrow (1977)
William J. Baumol (1977)
John G. Gurley (1977)
Anne O. Krueger (1977)
Carl Stevens (1977)
Thomas E. Weisskopf (1977)

COMMITTEE ON PUBLICATIONS
Michael Lovell, *Chair* (1977)
Edwin Burmeister (1979)
Peter Diamond (1979)
John G. Gurley (1978)
Robert Ferber (1978)
Robert Gallman (1978)
C. Elton Hinshaw, *ex officio*

COMMITTEE ON THE STATUS OF WOMEN IN THE ECONOMICS PROFESSION
Barbara Reagan, *Chair* (1977)
Mariam Chamberlain (1979)
Ann Friedlaender (1979)
William F. Hellmuth (1979)
Janice Madden (1978)
Isabel Sawhill (1977)
Margaret Simms (1978)
Nancy Teeters (1977)
Lawrence R. Klein, *ex officio*

FINANCE COMMITTEE
Robert Eisner, *Chair* (1977)
Robert G. Dederick (1979)
James Lorie (1978)
Rendigs Fels, *ex officio*

ECONOMICS INSTITUTE POLICY AND ADVISORY BOARD
Edwin S. Mills, *Chair* (1978)
Raymond Vernon (1980)
Carlos F. Diaz-Alejandro (1980)
Arnold Harberger (1977)
Richard H. Holton (1977)
Paul G. Clark (1978)
Carl Keith Eicher (1979)
Anne O. Krueger (1979)

U.S.-SOVIET EXCHANGES
Lloyd G. Reynolds, *Chair*
Abram Bergson
John Meyer
Rendigs Fels, *ex officio*

Special Committees

SEARCH COMMITTEE FOR EDITORS
James Tobin, *Chair*
Irma Adelman
Albert Ando
George R. Feiwel
Martin S. Feldstein
John G. Gurley
Michael C. Lovell
Bernard Saffran

AD HOC COMMITTEE ON FEDERAL FUNDING OF ECONOMIC REARCH (1977)
Zvi Griliches
Robert Solow
Marc Nerlove

AD HOC COMMITTEE TO REVIEW NEW PROPOSED STANDARD OCCUPATIONAL CLASSIFICATION SYSTEM
H. Gregg Lewis, *Chair*
Victor Fuchs
Margaret S. Gordon
Michael Piori
Sherwin Rosen

COMMITTEE ON COMPUTERIZATION
John R. Meyer, *Chair*

NOMINATING COMMITTEE (1977)
Walter Heller, *Chair*
Nancy S. Barrett

Huey J. Battle
David M. Gordon
Bert G. Hickman
Irving B. Kravis
Ralph W. Pfouts

NOMINATING COMMITTEE (1978)
Andrew F. Brimmer, *Chair*
Carolyn Shaw Bell
Peter A. Diamond
John G. Gurley
Robert M. Haveman
Bert G. Hickman
Vernon L. Smith

COMMITTEE ON ELECTIONS (1977)
Ben Bolch, *Chair*
Barbara Haskew
C. Elton Hinshaw, *ex officio*

NEW YORK CITY LOCAL ARRANGEMENTS COMMITTEE (1977)
Richard G. Davis, *Chair*
Richard C. Aspinwall
Peter Bakstansky
Alice J. Christensen
Roger D. Collons
Alastair I. Hunter-Henderson
Eleanor M. Johnson
Barbara MacPhee
Robert Mathieson
Robert A. Schwartz
Violet O. Sikes
Thomas W. Synnott
Ingo Walter

CHICAGO LOCAL ARRANGEMENTS COMMITTEE (1978)
Karl A. Scheld, *Chair*
Roby L. Sloan, *Co-Chair*
Eugene A. Birnbaum
Bert E. Elwert
Walter D. Fackler
William L. Helfers
Jim Lilly
Laurence J. Mauer
Richard S. Peterson
Violet O. Sikes
Barbara Weaver

Council and Other Representatives

AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
William Nordhaus (1979) Section on Social and Economic Sciences

AMERICAN COUNCIL OF LEARNED SOCIETIES
William Parker (1978)

AMERICAN POLITICAL SCIENCE ASSOCIATION-JOINT RESEARCH PROJECT ON CONFIDENTIALITY OF RESEARCH SOURCES
Gary Fromm

FEDERAL STATISTICS USERS CONFERENCE
Edward F. Denison (1979)

INTERNATIONAL ECONOMIC ASSOCIATION
Franco Modigliani (1981)

INTERSOCIETY COMMITTEE ON TRANSPORTATION
William Dodge

JOINT AD HOC COMMITTEE ON GOVERNMENT STATISTICS
Edward F. Denison
Gary Fromm

JOURNAL OF RESEARCH ON CONSUMER BEHAVIOR
Kelvin J. Lancaster (1978)

NATIONAL ARCHIVES ADVISORY COUNCIL–GENERAL SERVICES ADMINISTRATION
Robert Gallman (1978)

NATIONAL BUREAU OF ECONOMIC RESEARCH
Carl F. Christ (1978)

NATIONAL COUNCIL FOR ACCREDITATION OF TEACHER EDUCATION
Philip Saunders (1977)

SOCIAL SCIENCE RESEARCH COUNCIL
Guy Orcutt (1977)
Robert Eisner (1978)

Representatives of the Association on Various Occasions–1977

INAUGURATIONS

Max Sherman, West Texas State University
Barry L. Duman

Donald Arthur Webb, Centenary College of Louisiana
Tom S. Sale III

John Edwin Johns, Furman University
Matthew A. Stephenson

Mary Evelyn Blagg Huey, Texas Woman's University
Kendall P. Cochran

Richard Franklin Rosser, DePauw University
Bernerd Bogar

Henry Jefferson Copeland, College of Wooster
Lucille G. Ford

Jarvis Ernest Miller, Texas A&M University
W. James Truitt

FIFTIETH ANNIVERSARY CONVOCATION OF THE UNIVERSITY OF HOUSTON
Gaston V. Rimlinger

C. ELTON HINSHAW, *Secretary*

Report of the Treasurer for the Year Ending December 31, 1977

The financial position of the American Economic Association improved dramatically in 1976, when there was a surplus of $150 thousand compared to cumulative deficits during 1969–75 totalling nearly half a million. The improvement resulted from the restructuring of dues that went into effect at the beginning of 1976, cost-cutting, sales of the *Index of Economic Articles*, improvement in the stock market, and settlement of a claim against Richard D. Irwin, Inc. Another large surplus was realized in 1977, and a surplus of comparable size is anticipated for 1978. (See Tables 1 and 2.)

By the end of 1978, the net worth of the Association, which had fallen to a negative figure at the end of 1974, will still be on the low side in spite of the recent surpluses. On December 31, 1977, the net worth was on the order of $431 thousand. During the ensuing twelve months, the rise in the net worth will be smaller than the surplus shown in Table 1 would suggest, because a new edition of the *Handbook* (or *Directory*, as the 1974 edition was called) will be published at a cost estimated at $150 thousand. Each year $50 thousand is budgeted for the next edition. In the years when no *Handbook* is published, the rise in net worth is correspondingly greater than the surplus; in years when one is published, the net worth rises less than the surplus.

The net worth serves as a cushion that would enable the Association to continue its activities in the face of unexpected adverse changes. There is no particular rule governing the optimum size of the net worth, but as long as it is lower than the deficits of 1969–75, it can hardly be considered too large. Consequently, the surpluses of 1976–78 are to be welcomed, and some further rise after that period is desirable. Nevertheless, the time is approaching when large surpluses will no longer be appropriate. Since expenses will continue to rise, the transition from large to small surpluses can best be accomplished by smaller increases in dues and subscription rates than would otherwise be necessary.

In recent years, receipts from dues and subscriptions have constituted about 70 percent of total revenues. In the absence of changes in dues and subscription rates, receipts from this source could be expected to grow no more than 5 percent per year from increased numbers of members and subscribers, probably less. Other sources of revenue will probably grow at about the same rate of 5 percent or less. Sales of the *Index* will fall as the pace of publication of new volumes slacks off to one a year. Investment income may grow if surpluses continue, the stock market rises, or rates of return increase. Other sources of income will rise more or less in step with the general growth of the economy and inflation.

Since more than half the expenditures of the Association are for printing its publications, the growth of expenditures will depend heavily on changes in costs in the printing industry, which could differ markedly from the general rate of inflation. Aside from printing, the operations of the Association are labor intensive, so costs would be expected to rise more rapidly than revenues (in the absence of changes in dues and subscription rates) except that economists' salaries are likely to rise less rapidly than other wages and salaries. In addition, the growth in the number of members and subscribers will add to costs as well as revenues. However, marginal costs are well below the rates charged members and subscribers, so that growth in numbers on balance is financially advantageous.

A rather optimistic set of projections would be for a growth rate of expenditures of 6 percent per year, a growth rate of revenues of 4 percent per year, and a decline in the surplus of 2 percent of revenues

TABLE 1—AMERICAN ECONOMIC ASSOCIATION, ACCOUNTING BASIS, 1976–78
(Thousands of dollars)

	1976 12 months Actual	1977 12 months Actual	1977 Budget (3-18-77)	1978 Budget (3-6-78)
REVENUE				
Operating Income				
Dues and subscriptions	726	744	750	795
Advertising	76	77	80	81
JOE Subscriptions	21	19	21	20
Sales—miscellaneous	28	36	30	36
Sales—mailing list	35	34	35	35
Sales—*Index*	51	100	60	87
Annual meeting	29	—	—	—
Other income	41	32	41	20
Total operating income	1,007	1,042	1,017	1,074
Settlement of Claim against Publisher	42	—	—	—
Investment Income				
Interest and dividends	33	41	35	45
Real capital gains (losses)	(37)	19	39	8
Total investment income	(4)	60	74	53
TOTAL REVENUE	1,045	1,102	1,091	1,127
EXPENSES				
Publications				
American Economic Review	269	284	290	305
Journal of Economic Literature	283	297	295	322
Handbook	47	50	50	50
Job Openings for Economists	25	27	25	27
Index of Economic Articles	13	31	—	23
Total publications expense	637	690	660	727
Operating and Administrative				
Salaries	98	103	108	110
Rent	8	9	8	10
Miscellaneous	67	88	96	85
Committees	29	38	39	41
Annual meeting	15	6	—	6
Federal income taxes	39	13	—	15
Total operating and administrative	257	257	251	267
TOTAL EXPENSES	894	946	911	994
SURPLUS (DEFICIT)	150	156	180	133

or expenditures per year, i.e., about $20 thousand a year. In such a case, there would be no need to increase dues and subscriptions during the next five years, unless some major increase in activities were desired. A rather pessimistic set of projections would be for a 10 percent rise in expenditures per year (about the increase expected between 1977 and 1978), a growth in revenues of 2 percent per year, and a reduction in the surplus of $80 thousand a year. In this case, although no increase in dues and subscription rates would need to be put into effect soon, a substantial increase would be called for in 1980. For planning purposes, I recommend an intermediate assumption that the surplus will fall $50 thousand a year. This implies a need to raise dues on January 1, 1981.

All this is on the assumption that the Executive Committee will undertake no new large expenditures like the ones it un-

TABLE 2—AMERICAN ECONOMIC ASSOCIATION CASH BUDGET, 1978
(Thousands of dollars)

		1976 Actual	1977 Budget	1977 Actual	1978 Budget
1.	SURPLUS (DEFICIT), Accrual basis	150	180	156	133
	Plus noncash charges to accrual budget				
2.	Reserve for *Handbook*[a]	50	50	50	(100)
3.	Capital losses (gains)	37	(39)	(19)	(8)
4.	Depreciation	1	1	1	1
5.	Subtotal, surplus (deficit) adjusted to cash basis	239	192	188	26
	OTHER OPERATIONS AFFECTING CASH				
6.	Increase (decrease) in deferred income	41	20	(121)[d]	150
7.	Increase (decrease) in accounts payable, etc.	16	0	60	(60)
8.	Decrease (increase) in accounts receivable	(117)	118	67	60
9.	Decrease (increase) in prepaid expenses	(1)	—	2	0
10.	Cash receipts less disbursements from restricted funds[b]	(29)	(5)	5	(3)
11.	Subtotal, other operations affecting cash	(91)	133	13	147
12.	TOTAL CHANGES FROM OPERATIONS	148	325	201	173
	INVESTMENT-TYPE TRANSACTIONS				
13.	Decrease (increase) in inventory of *Index* volumes	(46)	(30)	15	(13)
14.	Sales (purchase) of office equipment	(1)	(1)	(2)	(1)
	CHANGES IN INVESTMENT ACCOUNTS:				
15.	(Interest and dividends)	(33)	(35)	(41)	(45)
16.	Transfers of cash from (to) investment accounts	(145)	(262)	(100)	(189)
17.	Custodian and investment counsel fees	3	3	3	3
18.	TOTAL INVESTMENT-TYPE TRANSACTIONS	(222)	(325)	(125)	(245)
19.	INCREASE (DECREASE) IN OPERATING CASH[c]	(74)	0	76	(72)

[a]The surplus, accrual basis, shown in line 1 is net of an annual addition of $50 thousand to the reserve for new editions of the *Handbook*. In 1978 a new edition will be published at an estimated cost of $150 thousand.

[b]Excludes Economics Institute.

[c]Excludes cash in investment accounts.

[d]Due to late billing for 1978 associated with computerization, there was a substantial decrease in deferred income in 1977 instead of the normal increase, but this is an accident of timing rather than a substantive change in cash flow and will result in an offsetting change in 1978.

dertook the last time the financial situation was favorable. The large deficits of the early 1970's, causing the net worth of the Association to fall below zero at the end of 1974, compelled the Executive Committee and the editors to make economies which they may now want to reconsider. If so, dues increases may be needed before 1981.

RENDIGS FELS, *Treasurer*

Report of the Finance Committee*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 30, 1977, with costs and market values as of that date. The total market value of the securities portfolio at year-end was $735,245. After adjustments for cash additions and withdrawals, including a sizable addition of $250,000 in Association funds for permanent investment, we estimate that the Association's investment portfolio (on a total return basis) experienced a modest loss of 1.7 percent during 1977.

The $735,245 total includes the funds remaining from a Special Grant that was made by the Ford Foundation in January of 1969 and subsequently commingled with the Association's account. As of December 30, 1977, the Association's portion of the aggregate account was $683,088 or 92.9 percent, and the Special Grant represented the remaining $52,157, or 7.1 percent of the total.

In line with the policy of the Finance Committee established two years ago, the Association's investment portfolio continues to hold a combination of both common stock and fixed-income investments. Along with the $250,000 addition to the portfolio, a program of some restructuring was carried out during the year. There were additions in a number of issues already held. Sales were made in Seven Up and Citicorp. There were new purchases in McDonalds, Cities Service, Halliburton, J. C. Penney, and Weyerhaeuser.

In terms of the portfolio's investment performance, the Committee can report that the 1.7 percent loss referred to above, while a disappointment, was a much better overall result than that experienced by the widely followed market averages. On the same total return basis, the Dow Jones Industrial Average lost 13.2 percent and the Standard and Poor's 500 Index lost 7.3 percent for calendar year 1977.

As we move into 1978, the Finance Committee is aware of uncertainties in domestic and international investment markets. While these uncertainties suggest caution, it is the Committee's view that many of them have already been discounted by the market. Equities are priced at nearly their lowest relationships relative to earnings, cash flow, and asset values in the past quarter-century. Consequently, while the securities' markets may exhibit some near-term weakness, the Committee believes that a meaningful exposure to equities in the Association's investment portfolio continues to be warranted during 1978.

At its December 1977 meeting, the Finance Committee felt it appropriate in setting overall policy to take into consideration substantial temporary investment funds which have built up outside of the attached permanent portfolio. With these other additional funds in mind, it was voted

*The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which follows.

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 30, 1977

	Value	Percent	Estimated Income
Cash Equivalents and Short-Term Securities	$214,634	29.2	$13,466
Medium-Term Securities	0	0.0	0
Long-Term Securities and Preferred Stocks	0	0.0	0
Convertible Securities	0	0.0	0
Equity Securities	520,611	70.8	19,680
Total	$735,245	100.0	$33,146

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 30, 1977

	Amount	Price	Value	Unit Cost	Total Cost	Estimated Income
CASH EQUIVALENTS AND SHORT-TERM SECURITIES (29.2 percent)						
CASH EQUIVALENTS (0–1 YEAR) (22.4 percent)						
Cash			$3		$3	
Stein Roe Cash Reserves, Inc.	165,065	1	164,787	1	164,787[a]	9,904
Subtotal Cash Equivalents			164,790		164,790	9,904
Other Short-Term Securities (1–5 Years) (6.8 percent)						
U.S. Treasury Notes (7.125 11/30/79)	25,000	100	24,961	100	25,008	1,781
U.S. Treasury Notes (7.125 11/15/80)	25,000	100	24,883	100	24,930	1,781
	50,000		49,844		49,938	3,562
Subtotal Other Short-Term Securities			49,844		49,938	3,562
Total Cash and Fixed Income Securities			214,634		214,728	13,466
EQUITY SECURITIES (70.8 percent)						
Utilities (6.1 percent)						
Central and Southwest	2,000	16	32,000	12	24,556[a]	2,520
Banks (5.5 percent)						
First Bank System	800	36	28,600	25	20,320[a]	1,280
Other Financial (4.9 percent)						
Alexander and Alexander	500	51	25,375	19	9,325[a]	700
Foods and Related (4.8 percent)						
Philip Morris	400	62	24,750	44	17,726	660
Merchandising (4.8 percent)						
Penney, J.C.	700	36	24,850	34	24,143	1,036
Paper and Forest Products (4.7 percent)						
Weyerhaeuser	900	27	24,638	26	23,495	720
Machinery and Construction (5.0 percent)						
Halliburton	400	65	26,050	63	25,310	400
Energy (17.7 percent)						
Cities Services	500	53	26,688	51	25,478	1,500
Continental Oil	800	30	24,000	19	15,580[a]	1,120
Gulf Oil	800	27	21,400	17	13,321	1,520
MAPCO	500	40	19,813	18	8,855	550
			91,901		63,234	4,690
Drugs and Medical (7.5 percent)						
Abbott Lab.	400	57	22,600	35	14,136	480
Merck	300	56	16,650	52	15,631[a]	510
			39,250		29,767	990
Electrical Products (6.6 percent)						
General Electric	690	50	34,415	36	24,536[a]	1,518
Computers (6.3 percent)						
IBM	120	274	32,821	111	13,325[a]	1,382
Broadcasting and Publishing (4.8 percent)						
CBS	500	50	24,875	37	18,662[a]	1,200
Miscellaneous (21.3 percent)						
Disney	639	40	25,560	20	12,928[a]	204
Eastman Kodak	600	51	30,676	65	38,911[a]	1,260
McDonalds	500	52	25,750	46	23,220[a]	100
Minnesota Mining and Mfg.	600	49	29,100	54	32,473[a]	1,020
			111,086		107,532	2,584
TOTAL EQUITY SECURITIES			520,611		401,931	19,680
TOTAL SECURITIES AND CASH			735,245		616,659	33,146

[a] More than one cost basis.

that the $250,000 which was added to the permanent fund be put into equities. The investment advisor was instructed to maintain an equity ratio in the permanent fund which would result in an equity ratio for the combined Association resources, permanent and temporary funds, of between one-half and two-thirds. Further, the Finance Committee removed its previous maturity restriction on the investment of nonequity funds, authorizing the investment advisor to extend maturities when appropriate to take advantage of changes in the yield curve, especially in the intermediate-term sector.

ROBERT EISNER, *Chair*

Auditors' Report

To the Executive Committee of
The American Economic Association:

We have examined the statement of assets and liabilities of THE AMERICAN ECONOMIC ASSOCIATION (a District of Columbia corporation, not for profit) as of December 31, 1977 and 1976, and the related statements of revenues and expenses, changes in general and restricted fund balances, and changes in assets and liabilities for the years then ended. Our examination was made in accordance with generally accepted auditing standards, and accordingly included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the accompanying financial statements present fairly the assets and liabilities of The American Economic Association as of December 31, 1977 and 1976, and its revenues and expenses, changes in fund balances, and the changes in its assets and liabilities for the years then ended, in conformity with generally accepted accounting principles consistently applied during the periods.

Arthur Andersen & Co.

Nashville, Tennessee
February 3, 1978

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF ASSETS AND LIABILITIES
DECEMBER 31, 1977 AND 1976

Assets	1977	1976
CASH	**$ 121,831**	**$ 46,373**
INVESTMENTS, at market (Notes 1 and 2):		
Temporary investments	269,673	405,763
Permanent investments	735,327	491,084
	1,005,000	**896,847**
ACCOUNTS RECEIVABLE:		
Advertising, back issues, etc.	103,085	76,845
Sales of *Index of Economic Articles*	—	51,303
Receivable from publisher	—	41,924
Allowance for doubtful accounts	(1,190)	(1,000)
	101,895	**169,072**
INVENTORY OF *Index of Economic Articles*, at cost	**31,072**	**46,098**
PREPAID EXPENSES	**3,583**	**5,405**
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of $6,992 in 1977 and $6,174 in 1976	**7,737**	**7,027**
Total Assets	**$1,271,118**	**$1,170,822**

Liabilities and Fund Balances	1977	1976
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	**$ 228,285**	**$ 168,015**
DEFERRED INCOME (Note 1):		
Life membership dues	65,412	68,034
Other membership dues	191,337	282,852
Subscriptions	144,668	169,967
JOE	9,525	11,049
	410,942	**531,902**
ACCRUAL FOR *Directory* (Note 1)	**150,000**	**100,000**
FUND BALANCES:		
Restricted (Note 4)	53,111	48,595
Add (deduct)–Unrecognized change in market value of investments	(2,419)	9,067
(Notes 1 and 3)	50,692	57,662
General	451,602	277,198
Add (deduct)—Unrecognized change in market value of investments (Notes 1 and 3)	(20,403)	36,045
General fund-net worth	431,199	313,243
Total fund balances	504,713	325,793
Add (deduct)—Unrecognized change in market value of investments (Notes 1 and 3)	(22,822)	45,112
Net fund balance	**481,891**	**370,905**
Total Liabilities and Fund Balances	**$1,271,118**	**$1,170,822**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF REVENUES AND EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1977 AND 1976

	1977	1976
REVENUES FROM DUES AND ACTIVITIES:		
Membership dues and subscriptions	$ 481,370	$ 468,444
Nonmember subscriptions	262,224	257,273
Job Openings for Economists subscriptions	19,325	21,034
Advertising	76,695	76,032
Sale of *Index of Economic Articles*	100,254	51,303
Sale of copies, republications and handbooks	36,212	27,664
Sale of mailing list	33,850	35,202
Annual meeting	–	29,064
Sundry	32,049	40,921
	1,041,979	1,006,937
SETTLEMENT OF CLAIM AGAINST PUBLISHER	–	41,924
INVESTMENT GAINS (LOSSES)—Note 2	60,373	(4,281)
Net Revenues	**1,102,352**	**1,044,580**
PUBLICATION EXPENSES:		
American Economic Review	221,025	210,962
Journal of Economic Literature	297,280	283,012
Papers and Proceedings	62,730	58,518
Directory publication (Note 1)	50,334	47,455
Job Openings for Economists	26,920	24,926
Index of Economic Articles	31,405	12,807
	689,694	637,680
OPERATING AND ADMINISTRATIVE EXPENSES:		
General and administrative		
Salaries	102,664	98,032
Rent	8,721	8,078
Other (Exhibit I)	88,420	67,026
Committees	37,975	28,639
Annual meeting	5,816	15,398
Provision for federal income taxes (Note 6)	13,000	39,352
	256,596	256,525
Total Expenses	**946,290**	**894,205**
REVENUES IN EXCESS OF EXPENSES	**$ 156,062**	**$ 150,375**

The accompanying notes to financial statements and Exhibit I are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN GENERAL FUND BALANCE FOR THE YEARS ENDED DECEMBER 31, 1977 AND 1976

	Total	Operations	Market Value Adjustments
Balance at January 1, 1976	**$110,535**	**$(122,280)**	**$232,815**
Add—market value adjustments resulting from inflation (Note 1)	16,288	–	16,288
Add—revenues in excess of expenses	150,375	150,375	–
Balance at December 31, 1976	**277,198**	**28,095**	**249,103**
Add—market value adjustments resulting from inflation (Note 1)	18,342	–	18,342
Add—revenues in excess of expenses	156,062	156,062	–
Balance at December 31, 1977	**$451,602**	**$ 184,157**	**$267,445**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1977

	Balance at January 1	Receipts	Disbursements	Allocation of Investment Gains (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$40,744	$ 2,101	-	$4,460	$47,305
The Alfred P. Sloan Foundation, Chase Manhattan Bank and Ford Foundation grants for increase of educational opportunities for minority students in economics	-	57,123	(56,524)	-	599
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	1,068	1,121	(2,189)	-	-
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	1,067	-	(451)	-	616
The Carnegie Foundation grant for the committee on the status of women in the economics profession	5	-	(5)	-	-
The National Science Foundation grant for support of a joint *U.S.-USSR* Symposium on the Economics of Technological Progress	-	706	(706)	-	-
The Minority scholarship fund for minority students applying for graduate work in economics	5,000	-	-	-	5,000
The Ford Foundation grant for development of a consortium on graduate studies in economics for minorities	-	-	(1,220)	-	(1,220)
Sundry	711	100	-	-	811
	$48,595	**$61,151**	**$(61,095)**	**$4,460**	**$53,111**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN RESTRICTED FUND BALANCES FOR THE YEAR ENDED DECEMBER 31, 1976

	Balance at January 1	Receipts	Disbursements	Allocation of Investment (Losses) (Note 4)	Balance at December 31
The Ford Foundation grant for Economics Institute's orientation program for foreign graduate students of economics	$102,004	$ 6,398	$ (62,367)	$(5,291)	$40,744
The Alfred P. Sloan Foundation, Chase Manhattan Bank and Ford Foundation grants for increase of educational opportunities for minority students in economics	30,753	32,593	(63,346)	–	–
Funds reserved by the Association for publication of revised editions of *Graduate Study in Economics*, a guide originally published with funds from a Ford Foundation grant	–	5,158	(4,090)	–	1,068
The Asia Foundation grant for Asian economists' membership dues to The American Economic Association and related travel expenses	1,442	–	(375)	–	1,067
The Carnegie Foundation grant for the committee on the status of women in the economics profession	4,866	–	(4,861)	–	5
The National Science Foundation grant for support of a joint *U.S.-USSR* Symposium on the Economics of Technological Progress	–	11,660	(11,660)	–	–
The Minority scholarship fund for minority students applying for graduate work in economics	–	5,000	–	–	5,000
Sundry	611	100	–	–	711
	$139,676	**$60,909**	**$(146,699)**	**$(5,291)**	**$48,595**

The accompanying notes to financial statements are an integral part of this statement.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF CHANGES IN ASSETS AND LIABILITIES FOR THE YEARS ENDED DECEMBER 31, 1977 AND 1976

	1977	1976
Cash, beginning of year	$ 46,373	$ 121,950
SOURCE (USE) OF FUNDS:		
Revenues in excess of expenses	156,062	150,375
Add noncash charges		
Depreciation	1,117	1,030
Directory publication (Note 1)	50,000	50,000
Market value adjustments (Note 1)	(19,267)	37,314
Funds provided by operations	187,912	238,719
(Increase) decrease in		
Receivables and prepaid expense	68,999	(117,914)
Inventory of *Index of Economic Articles*	15,026	(46,098)
Investments	(108,153)	(165,116)
Office furniture and equipment	(1,827)	(644)
Increase (decrease) in		
Accounts payable and accrued liabilities	60,270	15,879
Deferred income	(120,960)	40,704
Restricted funds	4,516	(91,081)
General fund, market value adjustment	18,342	16,288
Unrecognized change in market value of investments	(48,667)	33,686
Cash, end of year	$ 121,831	$ 46,373

The accompanying notes to financial statements are an integral part of this statement.

NOTES TO FINANCIAL STATEMENTS

(1) Significant Accounting Policies

Investments:

The association accounts for its investments on a market value basis. Under the method used by the Association to value investments, the change in market value of corporate stocks during the year, after adjusting for an inflation factor (5.9 percent in 1977 and 4.7 percent in 1976), is recognized in income over a three-year period. The change in market value of treasury bills, commercial paper, etc., is reflected currently in income. The changes in market value of investments are allocated to the general and restricted fund balances as appropriate.

Accrual for Directory:

Approximately every three to five years, the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was last published in 1974 and distributed at no cost to the membership. In order to match more properly the publishing cost of this directory with revenue from membership dues, the Association has provided $50,000 annually (since date of last publication) for estimated publishing costs which will reduce actual directory expense in the year of publication.

Deferred Income:

Revenue from membership dues and subscriptions to the various periodicals of the Association are deferred when received; these amounts are then recognized as income as publications are mailed to the members and subscribers.

Revenue from life membership dues is recognized over the estimated average life of these members.

(2) Investments and Investment Income

The following is a summary of investments held by the Association at December 31:

	1977		1976	
	Cost	Market	Cost	Market
Treasury bills, commercial paper, etc.	$484,389	$ 484,389	$549,081	$549,081
Corporate stocks	401,947	520,611	204,943	347,766
Total	**$886,336**	**$1,005,000**	**$754,024**	**$896,847**

Investment gains (losses) recognized in income for the years ended December 31, were as follows:

	1977	1976
Treasury bills, commercial paper, etc.		
Interest	$31,088	$ 23,219
Change in market value	–	–
	31,088	**23,219**
Corporate stocks		
Cash dividends	10,018	9,814
Increase (decline) in market value recognized (Note 3)	21,552	(46,702)
	31,570	**(36,888)**
Less Investment gains (losses) allocated to restricted fund (Note 4)	**2,285**	**(9,388)**
Investment gains (losses) included in income	**$60,373**	**$ (4,281)**

(3) Unrecognized Change in Market Value of Investments

As described more fully in Note 1, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affect 1977 and 1976 revenues, and the amount of these market value increases (declines) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change December 31	
	1977	1976	1978	1979	1977	1976
1974	$ –	$(83,715)	$ –	$ –	$ –	$ –
1975	28,913	28,913	–	–	–	28,914
1976	8,100	8,100	8,099	–	8,099	16,198
1977	(15,461)	–	(15,460)	(15,461)	(30,921)	–
	$ 21,552	**$(46,702)**	**$ (7,361)**	**$(15,461)**	**$(22,822)**	**$45,112**

Included in the above unrecognized changes as of December 31, are increases (declines) of ($2,419) and $9,067 in 1977 and 1976, respectively, which have been allocated to a restricted fund. The amounts allocated are based on the percentage of the Association's total stock portfolio owned by this restricted fund.

(4) Restricted Fund

In 1968, the Association entered into an agreement with the University of Colorado relating to the Ford Foundation grant for the Economics Institute which provides, among other things, that the Association invest a portion of the funds received and allocate any income and market value adjustments therefrom to the restricted fund. In accordance with this agreement, the following adjustments were allocated to the restricted fund:

	1977	1976
Net investment gains (losses) (Note 2)	$2,285	$(9,388)
Market value adjustments arising from inflation	2,175	4,097
	$4,460	**$(5,291)**

(5) Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was $14,550 and $14,625 for 1977 and 1976, respectively.

(6) The Association

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under section 501(c)(3) of the U.S. Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists.

The Association has been determined to be an organization which is not a private foundation.

EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENT OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1977 AND 1976

	1977	1976
Mailing list file maintenance and periodic mailing expenses	$28,075	$17,344
Accounting and legal	17,450	10,770
Office supplies	11,179	13,744
Postage	14,136	7,592
Dues and subscriptions	2,906	2,590
Telephone	3,199	3,685
Investment counsel and custodian fees	2,624	2,979
President and president-elect expenses	1,765	3,525
Travel and entertainment	1,955	566
Depreciation (straight-line method)	1,117	1,030
Uncollectible receivables	-	789
Currency exchange charges (credits)	693	(620)
Insurance and miscellaneous	3,321	3,032
	$88,420	**$67,026**

Report of the Managing Editor
American Economic Review

The number of manuscripts submitted in 1977 was 690, approximately the same as last year, but somewhat lower than in previous years. We printed 114 papers in all, 3 fewer than last year. The comparative statistics for the last twenty years are shown in Table 1. As of November 15, the backlog of accepted papers is 71, slightly larger than last year at the same time. Twenty-six papers will appear in the March 1978 issue, and the remainder in June and September 1978. A paper accepted now would be printed in either September or December 1978, depending on its length.

The backlog of unprocessed manuscripts has also been cut substantially. As of November 30, all (or nearly all) manuscripts first received before May 1, 1977 have been processed. The authors of these papers all have received some type of decision from my office. It is not possible, given our present resources, to cut this delay time down any further.

Number and Subject Matter of Submitted and Printed Manuscripts

As Table 2 indicates, we printed 114 regular papers this year, 50 main articles and 64 shorter papers or communications. The size of the *Review* has been maintained at 1,067 pages, comparable to 1,035 in 1976 and 1,068 in 1975.

Table 3 presents a distribution of manuscripts classified by subject matter. The most popular fields are microeconomics, labor, macro-economic theory, international economics, monetary theory, and welfare theory. This has not changed from prior years.

Administration

The processing of manuscripts appears to be going more smoothly this year, for two reasons. First the screening is causing fewer delays. We are sending manuscripts out in smaller batches and they are returned more rapidly. We are also using more individuals as screeners, as seen from the fifty-two names who are acknowledged below. Second, the Board of Editors has taken on a different function, and one that saves me a great deal of time. The Board traditionally served as a group of super referees, doing about one-third of the papers sent out to be read. This has now changed. Instead the members of the Board have agreed to serve in an appeals capacity, reading manuscripts whose authors challenge a referee's judgment. Not all challenges are sent to the Board, only the difficult ones. I am grateful to them for their cooperation. I will discuss with the Board the possibility of extending this effort to communications as well.

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1957–77

Year	Submitted	Published	Ratio of Published to Submitted
1957	215	40	.19
1958	242	46	.19
1959	279	48	.17
1960	276	46	.17
1961	305	47	.15
1962	273	46	.17
1963	329	46	.14
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17

TABLE 2—SUMMARY OF CONTENTS, 1976 AND 1977

	1976		1977	
	Number	Pages	Number	Pages
Articles	52	657	50	635
Shorter Papers, including Notes, Comments and Replies	65	296	64	334
Special Articles	4	8	1	13
Dissertations		23		23
Announcements and Notes		42		53
Index		9		9
TOTAL	121	1035	115	1067

Expenses—Printing and Mailing

The 1977 printing and mailing costs were some $5 thousand (or 4 percent) greater than in 1976, yet below the projected budget of $150 thousand. These costs would have been significantly higher but for a change in our vendors. We now send the typesetting work to one establishment and the printing to a second, with our office acting as the contracting middleman. This has already resulted in a significant saving on the four quarterly issues of the *Review*, and I expect further savings on the *Proceedings* issue.

The budgeted expenses shown in Table 5 require some explanation. Beginning January 1, my office will take over the editing of the *Proceedings* issue of the *Review*. The 1978 budgeted item of $210 thousand for printing and mailing includes $60 thousand for the *Proceedings* plus an estimated $150 thousand for the four quarterly issues of the *Review*.

TABLE 3—SUBJECT MATTER DISTRIBUTION OF SUBMITTED AND PUBLISHED MANUSCRIPTS IN 1977

	Submitted	Published
General Economics and General Equilibrium Theory	16	0
Micro-Economic Theory	116	19
Macro-Economic Theory	64	8
Welfare Theory and Social Choice	45	10
Economic History, History of Thought, Methodology	11	2
Economic Systems	20	0
Economic Growth, Development, Planning, Fluctuations	29	5
Economic Statistics and Quantitative Methods	28	11
Monetary and Financial Theory and Institutions	49	9
Fiscal Policy and Public Finance	42	3
International Economics	59	14
Administration, Business Finance	21	0
Industrial Organization	39	3
Agriculture, Natural Resources	25	2
Manpower, Labor, Population	97	19
Welfare Programs, Consumer Economics, Urban and Regional Economics	29	9
TOTAL	690	114

Board of Editors

Five members of the Board of Editors will complete their terms at the end of this year: Eugene Fama; Robert J. Gordon; James Melvin; William Nordhaus; Anna Schwartz. In addition, Stephen Resnick's term expired March 31, 1977. I am deeply grateful to them for their high professional standards, work, and cooperation.

Last spring the executive committee appointed the following new members to the Board: Albert Ando; Elizabeth Bailey; David Bradford; Frederic Scherer. In addition it appointed Jerome Stein and Martin Feldstein to second terms.

TABLE 4—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING IN 1977

	Copies Printed	Pages Net	Pages Gross	Cost Issue[b]	Cost Reprints[c]	Total
March	27,500	265	304	$38,979.56	$570.14	$38,409.42
June	27,500	279	304	34,920.77	530.66	34,390.11
September	27,434	276	304	36,438.93	602.16	35,836.77
December[a]	27,500	247	288	36,000.00	600.00	35,400.00
TOTAL	109,934	1,067	1,200	$146,339.26	$2,302.96	$144,036.30

[a]Estimate.
[b]Includes allocated cost of preparing mailing list.
[c]Credit resulting from charges to authors for additional reprints.

I wish to express my thanks to the continuing members of the Board of Editors: Irma Adelman; David Baron; Robert Barro; Laurits Christensen; David Laidler; Frank Stafford.

I shall submit the names of new members of the Board to the Executive Committee at its meeting in March 1978.

Acknowledgments

I should like to thank my associates for their cooperation and patience: Wilma St. John for her fine work as assistant editor; Deborah Franklin our editorial assistant; and Sandra Szelag our secretary.

The following graduate students worked for the *Review* as proofreaders and hunters of false proofs: George Briden, John Chilton, Karl Donenwirth, Marvin Goodfriend, Robert King, and Phillip Kott.

TABLE 5—ACTUAL AND BUDGETED EXPENDITURES, 1970–78

	Printing and Mailing	Office Expenses	Total
1970	$111,227	$36,336	$147,564
1971	120,120	43,524	163,644
1972	107,196	44,473	151,669
1973	117,873	49,121	166,994
1974	139,502	58,396	197,898
1975	129,476	63,372	192,848
1976	139,300	67,130	206,430
1977[a]	150,000	71,663	221,663
1977[b]	144,036	70,851	214,887
1978[c]	210,000	94,539	304,539

[a]Budget.
[b]Actual.
[c]Budget to include *Proceedings* issue.

The following served as editorial consultants in the screening of manuscripts:

W. J. Adams
J. W. Albrecht
A. R. Beckenstein
E. R. Berndt
F. E. Bloch
G. Borjas
K. Boyer
R. Braeutigam
B. W. Brown
G. Butters
R. Deb
A. Denzau
G. Dorman
L. Edwards
D. Epple
R. Falvey
A. Feldman
R. Feldman
M. T. Flaherty
R. Forsythe
D. Frey
H. L. Gabel
S. Garber
J. Geweke
R. Gilbert
G. S. Goldstein
J. Gordon
R. Gordon
W. H. Greene
M. Harris
M. Hashimoto
G. G. Hildebrandt
R. J. Hodrick
N. M. Kiefer
C. Lieberman
D. L. McNicol
N. Nishimizu
N. P. Obst
A. M. Over, Jr.
L. Papademos
J. D. Richardson
J. Roberts
H. Rockoff
H. S. Rosen
M. R. Rosenzweig
T. Russell
S. Shavell
J. B. Shoven
C. Stone III
A. D. Strickland
R. Wilder
A. Zelenitz

In addition to the members of the Board and the editorial consultants, I have sought and received the assistance of a large number of referees during the year. I wish to thank them for their cooperation

and high standards in reading and evaluating manuscripts. The following have assisted as referees:

B. Aghevli
N. Aitken
A. Alchian
P. Allen
J. Anderson
S. Arndt
R. Artle
C. Azariadis
C. Azzi
M. J. Bailey
M. N. Baily
B. Balassa
R. Baldwin
G. Ballentine
R. Barlow
J. Barron
Y. Barzel
L. Bassett
R. N. Batra
W. J. Baumol
V. J. Bawa
M. Beckmann
F. Bell
L. Benham
Y. Ben-Porath
B. Bergmann
A. Bergson
T. Bergstrom
R. Berner
E. Berndt
S. Berry
T. J. Bertrand
H. Binswanger
F. Black
A. Blinder
M. K. Block
M. Blume
R. Bodkin
J. Bonin
K. Borch
T. Borcherding
M. Boskin
K. Boyer
R. Boyer
W. Branson
F. P. R. Brechling
G. Brennan
A. Brillembourg
J. Brittain
M. Bronfenbrenner
J. P. Brown
M. Brown
R. Bryant
J. M. Buchanan
K. Burdett
E. Burmeister
G. Butters
P. Cagan
G. Cain
J. Callen
G. Calvo
T. F. Cargill
W. Carleton
J. A. Carlson
D. Carlton
R. Carson
G. Chamberlain
P. L. Cheng
S. Cheung
B. R. Chiswick
G. Chow
C. F. Christ
C. Clotfelter
P. R. P. Coelho
R. Coen
B. C. Conley
J. Conlisk
M. Connolly
J. C. Cox
J. Cox
R. L. Crouch
G. Daly
R. Dansby
M. R. Darby
R. d'Arge
E. Davis
K. Davis
R. H. Day
R. Deacon
A. Deardorff
G. De Menil
E. Denison
A. Denzau
A. De Vany
D. Dewees
P. A. Diamond
S. Diller
A. K. Dixit
P. Doeringer
F. T. Dolbear, Jr.
W. Dolde
M. P. Dooley
G. Dorman
R. Dornbusch
L. Dudley
R. Dusansky
C. Eaton
R. Ehrenberg
I. Ehrlich
R. Eisner
B. Ellickson
J. W. Elliott
T. W. Epps
R. Fair
L.-S. Fan
H. S. Farber
G. Faulhaber
E. Feige
G. Feiger
A. Feldman
P. J. Feldstein
R. Fernandez
G. S. Fields
R. Findlay
J. M. Finger
D. Fischer
S. Fischer
P. C. Fishburn
A. Fisher
A. Fishlow
R. Flanagan
B. Fleisher
D. Foley
E. Foster
R. Freeman
A. Friedlaender
B. Friedman
B. M. Friedman
J. Friedman
I. Friend
G. Fromm
E. Furubotn
M. Fuss
L. E. Gallaway
G. C. Galster
A. Gandolfi
I. Garfinkel
D. Gaskins, Jr.
F. Gehrels
H. Genberg
M. A. Ghali
L. Girton
C. Goetz
S. M. Goldfeld
F. L. Golladay
M. J. Gordon
J. P. Gould
H. Grabowski
D. A. Graham
H. N. Gram
E. Greenberg
M. Greenhut
P. Greenwood
J. Griffin
R. Gronau
H. Grossman
S. Grossman
J. Gwartney
W. J. Haley
R. Hall
R. Hamada
M. Hamburger
D. Hamermesh
B. W. Hamilton
H. Hansmann
E. A. Hanushek
J. Haring
J. P. Harkness
M. Harris
R. Hartman
J. M. Hartwick
J. Heckman
W. P. Heller
E. Helpmann
D. W. Henderson
J. V. Henderson
J. Hirschleifer
O. Hochman
W. Holahan
C. Holt
D. Holthausen
H. Hori
I. Horowitz
T. Horst
J. R. Hosek
D. Howard
P. Howitt
C. Hsiao

R. P. Inman
P. Isard
D. Jaffee
G. Johnson
L. Johnson
T. Johnson
R. Jones
R. W. Jones
M. Jones-Lee
P. Jonson
D. Jorgenson
P. Kalman
E. J. Kane
D. Katzner
T. Keeler
J. Kennan
M. W. Keran
J. Kesselman
A. Khan
M. Khan
R. Klein
P. R. Kleindorfer
A. K. Klevorick
R. Koenker
J. C. Koeune
T. Koizumi
M. E. Kreinin
M. Ladenson
R. J. Lampman
L. J. Lau
E. Lazear
A. Leijonhufvud
H. Leland
D. R. Lessard
D. Levhari
H. G. Lewis
C. Lieberman
L. A. Lillard
C. M. Lindsay
C. Link
N. Liviatan
C. Lloyd
C. B. Lloyd
D. Logue
J. Lothian
C. A. K. Lovell
R. E. B. Lucas
S. McCafferty
R. A. McCain
J. J. McCall
B. T. McCallum
H. McCulloch
R. McCulloch
L. McKenzie
J. McKie
R. McKinnon
J. MacKinnon
C. D. MacRae
W. Magat
S. Maital
J. H. Makin
G. Mandelker
M. Manove
M. Manser
J. Marchand
E. Marfisi
J. Marshall
C. Maurice
T. Mayer
W. Mayer
J. Mayshar
J. Medoff
A. Melnik
A. H. Meltzer
L. H. Meyer
P. Meyer
R. Meyer
M. Michaely
P. Mieszkowski
B. Miller
M. Miller
J. C. Miller III
H. Minsky
L. Mirman
E. Mishan
F. Mishkin
T. Miyao
H. Mohring
J. B. Moore
M. Morishima
J. R. Moroney
J. Muellbauer
A. H. Munnell
M. Mussa
R. Muth
J. Myers
E. Nadel
K. Nagatani
P. Neher
E. J. Nell
R. Nelson
E. Neuberger
G. Neuman
J. Newhouse
Y.-K. Ng
Y. Niho
R. Noll
W. Oakland
R. Oaxaca
E. Olsen
M. Olson
G. H. Orcutt
J. A. Ordover
L. Orr
D. K. Osborne
J. M. Ostroy
T. Page
M. Paglin
R. E. Park
M. Parkin
D. Parsons
P. Pashigian
D. Patinkin
M. V. Pauly
S. Peck
S. Pejovich
S. Peltzman
J. Pencavel
R. J. Penner
S. Perrakis
E. S. Phelps
M. Piore
J. E. Pippenger
C. Plott
C. G. Plourde
K. Polenske
R. Pollak
J. Pomery
W. Poole
R. B. Porter
R. Porter
R. A. Posner
D. Purvis
D. H. Pyle
J. Quigley
T. Rader
R. Radner
L. Rapping
R. H. Rasche
A. Raviv
A. Razin
U. E. Reinhardt
S. Resnick
M. Reynolds
J. Riley
S. Robinson
C. A. Rodriguez
R. Roll
D. Roper
H. Rose
S. Rose-Ackerman
S. Rosen
S. Rottenberg
J. C. R. Rowley
M. Rubinstein
W. R. Russell
J. Rutledge
E. Sadka
G. S. Sahota
D. Salkever
S. Salop
A. Sandmo
A. Santomero
T. Sargent
R. Sato
M. Satterthwaite
T. Saving
J. Scadding
S. Schaefer
L. D. Schall
F. M. Scherer
B. Schiller
R. Schmalensee
M. Scholes
R. Schuler
T. P. Schultz
W. Schulze
M. Schupack
R. Schwartz
W. Schwert
J. Scoville
G. W. Scully
J. Seater
A. K. Sen
R. S. Seneca
E. Sheshinski
J. Shoven
C. D. Siebert
J. Siegel
W. Silber
E. Silberberg
C. A. Sims
D. Sjoquist
G. Smith

K. Smith
L. B. Smith
V. L. Smith
R. Soligo
L. R. Southwick, Jr.
W. Springer
H. O. Stekler
B. Stigum
M. R. Straszheim
J. D. Stryker
W. J. Stull
D. Suits
J. Sweeney
P. J. Taubman
D. Taylor
J. Taylor
R. L. Teigen
P. Temin
M. Teubal
L. Thurow
N. Tideman
J. Tobin
R. Toikka
R. Townsend
W. Travis
E. Truman
H. Tuckman
S. Turnovsky
D. Usher
A. Vanags
J. Vanek
N. Van Long
H. R. Varian
E. C. H. Veendorp
W. S. Vickrey
D. R. Vining, Jr.
W. Vroman
M. L. Wachter
H. Wan
F. Warren-Boulton
H. W. Watts
R. Waud
W. E. Weber
L. Wegge
J. Weicher
R. L. Weil
B. Weisbrod
L. W. Weiss
Y. Weiss
F. Welch
R. Wertheimer
R. Westin
W. C. Wheaton
A. Whinston
L. J. White
C. R. Wichers
L. Wilde
S. Williamson
R. Willig
R. J. Willis
C. Wilson
D. Wise
K. Wolpin
S. Y. Wu
F. Wykoff
Y.-H. Yeh
W. P. Yohe
E. Zabel

GEORGE H. BORTS, *Managing Editor*

Report of the Managing Editor
Journal of Economic Literature

This annual report contains information both on the four issues of the 1977 volume of the *Journal of Economic Literature* (*JEL*) and on the appearance of the additional volumes of the *Index of Economic Articles* (there are now five, 1969–73, available for immediate purchase).

Table 1 illustrates the projected allocation of space in the *JEL* for 1977 as well as the comparisons for the years 1970–76. Table 2 classifies the material by subject matter both for the 1977 issue and the totals for the period 1969–77. And, finally, Table 3 classifies the material by technical difficulty.

Members will note that we published four survey articles during 1977, and that we published five essays on the literature or connected with the literature.

We have, in various stages of completion, commissioned survey articles on the present state of theory and known facts of income by share, the literature on the Phillips curve, the literature on the welfare implications of national accounting, and the literature on population superannuation. We have also in process essays on the reactions to the literature on the social discount rate, Malinvaud's new book on the theory of employment, the evolution of Sir John Hicks' economics, some recent work by Uzawa of the University of Tokyo, and Lord Kahn's recollections of the evolution of John Maynard Keynes's thinking about the General Theory.

During 1977 we have managed to bring to press *Annual Indexes* for the years 1972 and 1973. Volumes for 1969–71 came out one year earlier. These volumes are processed only in part as a by-product of the quarterly *JEL*. They also contain entries coming from *Festschriften* and collected essays. The amount of checking that is necessary to produce each of these volumes is virtually astronomical. As the members know, all of this material is processed by computer. Anyone noting an error in the Subject Index section of the quarterly *JEL* with regard to spelling or classification ought to notify me as soon as possible so that a correction can be made for the annual *Index*. We will publish the *Index* for 1974 in two or three months and the *Index* for 1975 will be in press by the autumn of 1978.

We have increased somewhat the number of journals which we list in the quarterly index of articles. Obviously anything listed there will also appear in the annual *Index*. Members may ask why we are increasing the list, particularly in this period of cost inflation, but the answer is simple. There are more journals being published than was previously the case.

The Chancellor and the Dean of the Faculty of Arts and Sciences of the University of Pittsburgh have again this year allocated some University of Pittsburgh support to the *JEL*. Their willingness to do so, particularly in this period of retrenchment and tight budgets, illustrates an understanding of and a devotion to scholarly work in the economics discipline. I regularly thank them privately and take this opportunity again to do so publicly. I should add, however, that it is illusory for the Association to believe that the University will be able to maintain indefinitely even the present level of support. It has already reduced support both in terms of student help and standard library assistance.

Four members of the Board of Editors have completed their terms. I wish to convey to them publicly (although I have already done so privately) my great appreciation of their tremendous help. Moses Abramovitz, Stanford University; Arthur S. Goldberger, University of Wisconsin; Thomas Mayer, University of California-Davis; and Ryuzo Sato, Brown University have been of inestimable help. No managing editor could ask for better associates than these. The other members of my

TABLE 1—QUANTITATIVE ANALYSIS OF *JEL* CONTENTS, 1973–77
(Number of pages in parentheses)

	1973		1974		1975		1976		1977	
	No.	Pages	No.	Pages	No.	Pages	No.	Pages	No.	Pages
Survey articles	4	(144)	3	(102)	3	(119)	3	(116)	4	(127)
Essays on subfields	2	(28)	5	(99)	5	(100)	8	(157)	4	(99)
Review articles	–	–	1	(5)	–	–	–	–	1	(9)
Articles about economic literature	2	(38)	–	–	1	(11)	–	–	–	–
Communications	10	(26)	13	(71)	12	(36)	2	(7)	15	(62)
Books annotated	1214	(239)	1211	(229)	1203	(223)	1204	(253)	1212	(246)
Books reviewed	175	(259)	168	(239)	183	(282)	185	(278)	172	(274)
Journal issues listed and indexed	1011	(185)	986	(180)	908	(177)	901	(159)	970	(174)
Number of individual articles	7218	–	7360	–	6788	–	6211	–	7164	–
Subject index of journal articles	–	(357)	–	(338)	–	(349)	–	(328)	–	(329)
Abstracts of articles	1906	(407)	1645	(312)	1637	(331)	1502	(309)	1589	(326)
Total pages[a]		(1748)		(1671)		(1700)		(1664)		(1713)

[a]Includes, in addition to listed pages, classification systems, table of contents, indices, journal subscription information, etc.

TABLE 2—CLASSIFICATION BY SUBJECT, 1969–77

		1977		1969–77
		Commissioned Survey	Creative Curmudgeon Essays	All Articles Total[a]
01	General	–	–	6
02	Theory	1	2	23
03	Thought (Methodology)	–	–	21
04	Economic History	–	–	3
05	Comparative Systems	–	–	4
11-12	Growth and Development	–	–	6
13	Stabilization	1	–	2
21-22	Econometric, Statistical Theory, Statistics	–	–	3
31	Monetary Economics	1	2	7
32	Fiscal Economics	1	–	5
40-44	International Economics	–	–	11
50	Managerial Economics	–	–	1
60	Industrial Organization, Industrial Regulation	–	–	1
70	Agricultural and Resource Economics	–	–	2
80	Labor Economics	–	–	6
90	Applied Welfare Economics, Regional Economics	–	1	7
	TOTALS	4	5	108

[a]Includes all review articles on books, general essays on all literature.

TABLE 3—CLASSIFICATION BY TECHNICAL DIFFICULTY, 1969–77

	1977		1969–77 Totals: Surveys Creative Curmudgeon Articles, Others[a]
	Surveys	Creative Curmudgeon Articles	
Most Difficult	–	1	20
Some Difficulty	4	1	50
Not Difficult	–	3	38
TOTALS	4	5	108

[a]Review articles or books and general essays on all literature; excludes very short communications.

Board have also been simply superb. I look forward to my continuing association with them. I have nominated several people to replace the departing four.

I also wish to thank the following (plus three who have chosen to remain anonymous) for advice and assistance on the commissioning, refereeing, and revising of articles:

Carlos Diaz-Alejandro
Kenneth J. Arrow
Anthony Atkinson
Leslie Barnett
William Baumol
Gerard Butters
Phyllis Deane
Herbert Giersch
H. Scott Gordon
Hendrik S. Houthakker
Anne Krueger
David Laidler
Robert J. Lampman
Allan Meltzer
Hyman P. Minsky
Franco Modigliani
Alan Peacock
Morris Perlman
Alan Prest
Reuben Slesinger
Carl Taylor
Lester G. Telser

Finally, the *JEL* staff, in its offices in Pittsburgh, has done magnificent work during this year. We have managed to bring out our issues more or less on schedule (the delays are in fact caused by the printer) while at the same time preparing several Indexes for publication. The associate editor, Naomi Perlman, has done, besides her own work, most of the supervisory work during a good portion of the year while I was on sabbatical at the University of Cambridge in England. The assistant editor, Drucilla Ekwurzel, combines the best of the world of cost-effective copy editing with considerable talent as a professional economist. We have also had considerable help from Lyndis Rankin (the principal secretary) and from Margaret Yanchosek (who handles the record keeping involved in the indexing process), and from a variety of other personnel who have worked full or part time for the *Journal*.

MARK PERLMAN, *Managing Editor*

Report of the Director
Job Openings for Economists

During 1977, employers advertised 1,470 new vacancies. Of these, 1,000 (68 percent) were classified as academic and 470 (32 percent) were nonacademic. Last year, employers advertised 1,265 new vacancies; 68 percent were academic and 32 percent were nonacademic. The division between academic and nonacademic remained the same, but the total number of new vacancies increased by 16 percent. Table 1 shows the total listings (employers), total vacancies advertised, new listings and new vacancies by type for each issue of *JOE* in 1977.

TABLE 1—JOB LISTINGS FOR 1977

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	101	193	76	132
April	88	143	74	117
June	39	70	36	62
August	40	91	38	85
October	77	185	70	156
November	96	215	96	215
December	149	362	102	233
Subtotals	590	1,259	492	1,000
Nonacademic				
February	22	81	14	45
April	21	66	18	51
June	24	52	22	47
August	22	80	19	71
October	25	106	20	77
November	17	58	17	58
December	28	138	23	121
Subtotals	159	581	133	470
TOTALS	749	1,840	625	1,470

Universities with graduate programs and four-year colleges continue to be the major source of job listings. They constituted 46 and 32 percent, respectively, of total employers. This is comparable to last year's 50 and 31 percent for the two. Table 2 shows the number of employers by type for each 1977 issue. The distribution is similar to that in 1975 and 1976.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economists that employers are seeking. The applied area of specialty seems to be of secondary importance. Table 3 shows the number of citations by field classification during 1977. General economic theory (000) led, followed by monetary and fiscal (300), econometrics and statistics (200), welfare and urban (900), and industrial organization (600).

Pursuant to the request of many members, a special supplementary issue was issued in November. Only new vacancies were listed. The issue was second only to the December issue in the number of new jobs advertised (see Table 1). The experiment was successful. We

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1977

Issue	Four-Year Colleges	Universities with Graduate Programs	Junior Colleges	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	42	59	–	6	–	2	1	10	3	123
April	51	37	–	3	6	–	–	9	3	109
June	20	19	–	6	1	3	3	11	–	63
August	10	30	–	5	1	2	–	11	3	62
October	25	52	–	8	3	2	1	9	2	102
November	37	59	–	3	2	4	1	6	1	113
December	58	91	–	8	5	3	2	8	2	177
TOTALS	243	347	–	39	18	16	8	64	14	749

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1977

Field[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	98	83	39	38	80	99	174	611
Growth and Development (100)	30	22	14	19	23	23	47	178
Econometrics and Statistics (200)	37	35	19	21	34	39	68	253
Monetary and Fiscal (300)	29	39	20	23	45	47	90	293
International Economics (400)	19	22	6	20	22	30	43	162
Business Administration, Finance, Marketing and Accounting (500)	37	42	10	12	29	31	54	215
Industrial Organization (600)	30	26	13	22	36	40	65	232
Agriculture and Natural Resources (700)	25	20	10	14	29	23	42	163
Labor (800)	24	12	17	14	19	30	43	159
Welfare and Urban (900)	34	31	12	14	35	46	71	243
Related Disciplines (A00)	7	9	3	3	5	5	8	40
Administrative Positions (B00)	10	9	4	10	5	6	6	50
TOTALS	380	350	167	210	362	419	711	2,599

[a]Fields of specialization codes are from the *Journal of Economic Literature*.

now plan to continue issuing a supplementary issue in November.

The proposed 1978 budget and the 1977 (adopted and actual) and 1976 (adopted and actual) budgets are given in Table 4. The 1977 approved budget for *JOE* projected a deficit (including allocated costs) of $4 thousand. The estimated actual deficit is $3 thousand. Total revenues are expected to be $22,500, total direct costs $11 thousand, and total indirect costs $14,300. If indirect costs are excluded, *JOE* continues to be self-financing. Our auditors have opined that *JOE* is an "unrelated business" and profits are taxable. The application of commonly accepted accounting principles

TABLE 4—*Job Openings for Economists*
Budget for 1978 (in thousands)

	1978 (Proposed)	1977 (Adopted)	1977 (Estimated Actual)	1976 (Adopted)	1976 (Actual)
Revenue					
Subscriptions			22.3		
Miscellaneous			.2		
Total Revenue	$21	$21	$22.5	$21	$21
Expenses					
Direct					
Computer	1		.8		
Typewriter Rental	2		2.4		
Postage	3		3.4		
Printing	4		3.6		
Salaries	–		.2		
Miscellaneous	1		.7		
Total Direct	11		11.1		
Indirect					
Salaries	15		14.3		
Total Expenses	$26	$25	$25.4	$15	$25
SURPLUS (DEFICIT)	(5)	(4)	(2.9)	(5)	(4)

should be sufficient to keep *JOE* from showing a profit in 1978 and subsequent years. The proposed budget for *JOE* for next year projects revenues of $21 thousand, total direct costs of $11 thousand, and total indirect costs of $15 thousand. This leads to a projected accounting deficit of $5 thousand.

JOE is virtually a one-woman operation. Violet O. Sikes is listed as coordinator of *JOE*, a title that is not very descriptive of her role. She handles the typing, editing, layout, subscriptions, and oversees the production. I wish to express my great appreciation for her dedication and hard work.

C. ELTON HINSHAW, *Director*

The Committee on The Status of Women in the Economics Profession

A major concern of the Committee on the Status of Women in the Economics Profession in 1977 was the need to increase and ensure opportunities for participation by women economists in the annual meetings of the American Economic Association. Such participation includes organizing and chairing sessions, presenting papers, and giving formal discussions of papers. An important part of the program process is the preplanned publication in the *Proceedings* issue of the *American Economic Review*.

By custom, the President-elect of the Association plans the overall program. Normally he selects a theme(s) for that year, selects the chairs for the Association sessions, approves the number of sessions the Association jointly sponsors with other members of the Allied Social Science Association, and has varying degrees of input on the selection of chairs for the joint sessions. He may or may not set guidelines or have informal requests which he makes of the session chairs. He may or may not have a program committee. In conjunction with the editor of the *Proceedings* issue of the journal, he decides which sessions are to be promised publication. This is an important incentive and bonus. In the case of at least one standing committee (not CSWEP), the Executive Committee of the Association has voted a policy of promising publication of papers from sessions to be planned by that committee for several years in the future, with an option to renew the policy at the end of the period.

In the six years of its existence,[1] the Committee has worked with each of the presidents-elect in turn to encourage them to ask their designated chairs to open up the informal network to include women economists. In addition, in each of the six years, the Committee has been asked by the President-elect to sponsor a session at the annual meeting. The first three years, the program dealt with ways to obviate sex discrimination in the economics profession. In the last three years the Committee has sponsored programs fitting in with the president-elect's topical themes. This year the topic was Macroeconomic Goals and Changing Labor Force Participation of Women. Next year the topic will be Equity: Individual versus Family. In planning these programs we have had two goals besides compatibility with the overall program themes. One has been to encourage increased participation by women economists. The other has been to encourage research by female or male economists on economic topics related to women. We have had both men and women economists on the programs in all six years, although women have predominated. We have been offered full publication rights in all but one year; at that time we were given only partial publication of the papers. On behalf of the Committee, I want to thank the previous presidents of the Association for their helpfulness in these matters. Without this, the participation level of women economists would have been far less.

The presidents-elect have varied enormously in the number of sessions sponsored, the number of formal papers versus round table discussions, the extent of participation with other groups in the Allied Social Science Association in joint sessions, and in their explicit concern with opening up the dominant informal networks to less well-known economists, men or women. In some cases, explicit requests to the program chairs to diversify the group giving papers have resulted in negligible pattern changes.

This year, to try to help further increased participation by women, the Committee has done two things. We have started a card

[1] The Committee was established in the spring of 1972 following the affirmative action resolutions encouraging women to participate in the economics profession passed by the Association at the New Orleans meeting in December 1971.

index file of current research by women economists, which we hope to develop in the future into a viable resource for program chairs. Second, we have made a statistical summary by sex of Association programs and publication of program papers since 1969.[2]

As a rule of thumb in interpreting the statistical summary, I would urge that a miniumum of 10–15 percent of the program participants should be women since (1) the annual proportion of women among those receiving Ph.D.'s averaged 11 percent from 1971–72 through 1975–76, and (2) the 1970 Census showed women comprised 14 percent of economists teaching at colleges and universities. In a spirit of affirmative action to redress previous imbalance, a goal of 15 to 20 percent women would be reasonable. The Committee now has on its computerized roster over 1,900 women economists. Of these, about 750 have Ph.D.s in economics. This pool should more than adequately support a goal of 15 to 20 percent women participants. Using the 15 percent goal for 1977, for example, would have translated into 12 women as session chairs, 48 as author or joint author of papers, and 22 as discussants. The program for 1977 had 80 sessions, far more than any program in the last nine years. The previous year had 50 sessions. The 15 percent goal for 1976 would have translated into 8 women as session chairs, 24 women as authors or joint authors, and 16 women as discussants. In actuality, instead of 15 to 20 percent, the participation of women in 1976 and 1977 were 14 and 6 percent, respectively, as session chairs, 12 and 8 percent as paper authors, and 7 and 9 percent as discussants.

Another way of looking at the data is to consider whether there has been any appreciable improvement of women economists' participation since the adoption of the affirmative action resolutions by AEA in December 1971. For this purpose, the three years preceding formation of CSWEP can be compared with the six subsequent years. In the later years, the Committee session alone adds 4 to 6 women. Hopefully, increased awareness of other chair persons should add considerably more. Data on which Appendix Tables 1–3 are based are summarized below in Table 1.

TABLE 1

	1969–71 Number Women per Year	1969–71 Percent of Total	1972–77 Number Women per Year	1972–77 Percent of Total
AEA Sponsored Sessions:				
Session chairs	1.7	7.3	2.7	11.4
Authors of papers	2.7	4.8	10.2	13.7
Discussants	3.3	4.7	6.3	11.6
Joint Sessions:				
Session chairs	.3	3.3	1.7	6.5
Authors of papers	3.7	3.2	7.3	6.7
Discussants	1.0	4.2	5.6	8.6

In total the number of times a woman appeared on the AEA sponsored program as session chair, paper author or joint author, or discussant increased from nearly 8 per year in 1969–71 to 19 in 1972–77. In joint sessions, the number increased from 5 per year in 1969–71 to 15 in 1972–77. As is true of men participants as well, these numbers represent even fewer individual women because of multiple appearances such as chair of one session and paper author at another. Trying to diversify and avoid excessive multiple appearances are perennial problems for program planners. It suggests that more centralized planning of the program could be useful.

In terms of both numbers and proportions, the opportunities for women to participate in the annual Association program have increased in the last six years. In the last three years although the proportions have not changed much, the numbers of women as authors or joint authors have been enhanced by the increased number of sessions. (See Appendix Table 1.)

In these tabulations, single authors of papers and multiple authors of papers were

[2]Thanks are given to Patricia Kirby Cantrell for help with the tabulations.

given equal weight on the basis that the important element for career advancement is to be on the program. Whether this method gives different results from a method where joint authors are considered to be .5 or .3 of an author, depends on whether female economists tend to be joint authors more than male economists do and whether the proportion of joint authors has changed over time. Both are researchable questions.

Considering all sessions sponsored by the Association, either alone or jointly, women are slightly more apt to be joint authors than are men. In addition, the trend over the last nine years, especially for men, has been to have more multiple author papers. Multiple authorship from 1969 to 1977 in all sessions at the annual Association meetings (except presidential addresses and special lectures) is shown in Table 2.

TABLE 2

Year	Female Coauthors as Percent of All Female Authors on Programs	Male Coauthors as Percent of All Male Authors on Programs
1969–71	40.0	29.5
1972–77	45.0	40.8

It should be noted that opportunities for women economists to participate in the annual sessions sponsored alone by the Association have been greater than in the sessions it sponsors jointly with other members of the Allied Social Science Association. Future presidents-elect of the Association may be able to give some leadership to increasing opportunities for women economists in the joint sessions.

One other major aspect of participation by women economists in the annual meetings is the opportunity to have their papers or discussions, when they are asked to be on the program, published in the *Proceedings* issue of the *AER*. Numbers of authors or multiple authors whose papers or discussions were published by AEA are shown in Table 3.

TABLE 3

Year	Number of Authors per Year		Number of Discussants per Year	
	Female	Male	Female	Male
1969–71	2.7	60.0	1.3	33.3
1972–76	5.4	73.4	.8	8.4

Because publication by and large is promised in advance by the president-elect of the Association in his capacity as overall program chair, the sessions in which publication is promised tend to be the more

TABLE 4—PUBLISHED PAPERS AND DISCUSSIONS FROM ANNUAL PROGRAM, BY SEX, 1969–76[a]

Year[b]	Number of Authors	Female Authors		Number of Discussants	Female Discussants	
		Number	Percent		Number	Percent
1969	65	1	1.5	56	2	3.6
1970	58	3	5.2	38	1	2.6
1971	65	4	6.2	10	1	10.0
1972	70	2	2.8	11	1	9.1
1973	87	5	5.7	9	1	11.1
1974	67	9	13.4	13	2	15.4
1975	87	5	5.7	7	0	0
1976	83	6	7.2	6	0	0

[a]Published in *American Economic Review Proceedings* (excludes presidential addresses and special lectures).

[b]Year of meeting. The *Proceedings* are published in the following year, usually in May.

prestigious sessions. The Committee has worked very hard on the issue of promised publication for CSWEP-sponsored sessions, and for most years has been successful in dealing with individual presidents-elect. Each year, however, is a new ball game. Unfortunately, the increase in women's papers published shown above is largely due to the Committee-sponsored sessions. Again, we are most appreciative of the Association presidents who have offered us this privilege. There is considerable room for improvement in the number of women economists asked to participate in the sessions preordained for publication, as shown by the numbers above and the percentages in Table 4.

I want to thank the members of our six-person committee[3] who have worked so hard this year to carry out the mandate of the Association to a) support and facilitate equality of opportunity for women economists in all aspects of economists' professional activities and b) help eradicate any institutional or personal discrimination against women economists. The commitment of the Association to these purposes is shown by the fact that this is the fourth year since CSWEP was designated a standing committee of the Association, and by its financial support of our basic activities.

[3]Membership from March 1972 to date has included: Walter Adams, Michigan State University; Carolyn Shaw Bell, Wellesley College (Chair, 1972 and 1973); Francine Blau, University of Illinois; Martha Blaxall, Health, Education and Welfare; Kenneth E. Boulding, University of Colorado; Mariam Chamberlain, Ford Foundation; Ann F. Friedlaender, Massachusetts Institute of Technology; John Kenneth Galbraith, Harvard University; Walter W. Heller, University of Minnesota; Janice Madden, University of Pennsylvania; Collette Moser, Michigan State University; Barbara B. Reagan, Southern Methodist University (Chair, 1974–77); Isabel Sawhill, Urban Institute; Margaret Simms, Atlanta University; Myra Strober, Stanford University; Nancy Teeters, Budget Committee, House of Representatives; Phyllis Wallace, Sloan School, Massachusetts Institute of Technology; Florence Weiss, National Economic Research Associates, New York City. In addition, the current president of the Association served *ex officio*. Our apologies to the two past presidents serving regular committee memberships, as well as *ex officio*, whose names were removed by a proofreader from the 1976 report.

During the year the Committee has worked to improve the operation of the market for economists, to increase the supply of women economists, and to add to the research information on the status of women economists. We have also encouraged economic analysis of public policies which affect all women, including women economists. We feel that it is imperative that we collect and analyze data as a basis for our policy recommendations. The activities of the year summarized below support one or more of the above Committee goals.

I. Roster

We have again this year updated the data for each woman economist on our computerized roster by sending each a copy of the previous material she supplied us on areas of specialization, highest degree in economics, school of highest degree, current professional rank or grade, current employer, address, and availability for new employment. We added new members and lost some, with the final number approximately the same as last year.

Prospective employers who requested the service were supplied with a subset of women economists who meet the criteria specified in the request. The prospective employers are then free to contact the women listed to ascertain whether there is mutual interest in the job match. Use of this service, which is made available at a nominal charge, continues to grow.

II. CSWEP Newsletter

Three issues of the CSWEP *Newsletter*, fall, winter, and spring, have been sent to all women economists on our roster. The fall issue was also sent to department chairs in the Chairman's Group and to Association officers. The *Newsletter* gives information of special interest to women economists, summarizes Committee activities, calls for abstracts of paper proposals for the annual AEA meetings, lists conferences and program plans for regional economics meetings, lists grant or fellow-

ship opportunities, notes research findings or publications of special interest, and presents short items submitted by individual women members of the Association.

This year we have also used the CSWEP *Newsletter* to request payment of $3 dues to become an associate member of CSWEP. These dues are in addition to the regular dues paid directly to the Nashville office of the Association.

The CSWEP *Newsletter* is sent bulk mail to reduce mailing costs. This often delays delivery. In spite of the delay, a survey made this year showed that the *Newsletter* is a popular and greatly appreciated service of the Committee. Our associate members want to see the *Newsletter* strengthened, but not abandoned. It clearly has been one of our major techniques to build an informal network among women economists across the country.

The *Newsletter* also carries a section of brief announcements of job openings for economists. The section is made up of those written notices which are sent us by the employers. The marginal cost of carrying these job notices is low, and no charge is made for the service.

The Committee recently completed an extensive evaluation of the usefulness of this service, and found that it is considered valuable by many women economists and many employers. The job listings only partially duplicate the jobs listed in *Job Opportunities for Economists* (*JOE*), and the *Newsletter* carries a note to remind women economists actively in the job market to also subscribe to *JOE*. The Committee has decided to continue to carry job listings for the immediate future as a further effort to improve the job market information flow.

III. National and Regional Meetings

At the annual Association meeting in New York City, CSWEP kept a hospitality room open and staffed with a committee member and volunteer associate members for two and a half days. Although the location this year was less than central, women economists and a few employers found their way to it. An extensive list of job opportunities received since the October *Newsletter* went to press was mimeographed, and distributed at the CSWEP room.

The program session sponsored by the Committee, mentioned early in the report, was well attended by men and women economists. Discussion was lively and extended in spite of the session being scheduled at the end of the meeting.

The Committee also sponsored an open meeting on the first day of the sessions. Although numerous topics were discussed by the associate members, the liveliest topic was the concern expressed by members from various parts of the country that many Association members, men as well as women, may not want to attend meetings in Chicago or Atlanta in 1978 and 1979 if Illinois and Georgia do not ratify the ERA.

As an experiment this year, the Committee cosponsored a special program session at the Southern Economics Association meetings. Papers and discussion centered on economic aspects of at-home time. The Committee also had a booth in the exhibit section at the Southern meetings, with an opportunity for women economists to register for our roster. A special letter was sent to each woman economist living in the southern quadrant of the United States urging them to come to the SEA meetings and advising them of the Committee's participation. The experience with the SEA suggests that continuation and expansion into other regional economic sessions may be a useful way to strengthen our services to women economists.

IV. Research on Salaries of Economists

In 1975 data were collected by the Committee on education and career patterns and current salaries of 710 women economists and from a paired sample of more than 1,200 male and female economists who did their academic work for their highest degree in economics at the same university at the same time. An econometric analysis of the factors influencing the income dif-

ferences between the men and women in spite of their similar investment in human capital was completed this fall by Myra Strober and Barbara B. Reagan. The first draft of the prospective report is now being reviewed. This research gives particular attention to the effect of gaps in women's work history, and finds that relatively few women economists have had such gaps, that those who did have gaps indeed incurred a salary penalty, but that sex per se is a far more important variable than gaps in employment in explaining income variation. This unusually rich data source for a relatively homogeneous group of professional workers permits an extensive list of variables to be considered. Some of the variables, notably the gaps in work history and number of times moved to accommodate a spouse's job needs, are not often available.

V. Academic Labor Market, 1975-76[4]

Women represented about the same proportion of the Ph.D. degree recipients in 1975-76 as the previous year, about 10.5 percent (see Table 5).[5] The number of Ph.D. degrees awarded per department reporting was up slightly. Departments in the Chairman's Group, sometimes called the Cartel, awarded 11.2 Ph.D. degrees in 1975-76 per department reporting compared with 10.5 in 1974-75 per department reporting. The other departments awarding Ph.D. degrees reported 4.7 Ph.D. degrees awarded in 1975-76 compared with 3.8 per department reporting the previous year.

In contrast, the proportion of women earning M.A. degrees in economics in 1975-76 was 13 percent, less than the previous year's 18 percent. Similarly, women receiving bachelor level degrees in economics was less than the previous year, 18 percent compared with 22 percent last year. Informal checks with faculty members in several different areas of the country suggest that increased interest by business in hiring women economists, particularly at entry levels, has attracted increasing numbers of women into business majors. Some of this increase is probably attracting away some of the women who otherwise might have chosen economics at the bachelors or masters level, and well may be even reducing the number of women choosing to get a Ph.D. in economics. Men have long been aware of business opportunities with payoffs as great or greater for an MBA as for a Ph.D. in economics. Women are now beginning to feel more welcome in business, and hope for movement up the career ladder in substantially new ways in large business enterprises.

About 75 percent of all Ph.D. students in economics in the fall of 1976, men and women, received financial aid—tuition, stipend, or both (Table 6). Nearly 40 percent of the M.A. students in economics also received financial aid. At the Ph.D. level the proportion receiving financial aid was the same as the previous year. At the M.A. level, however, the proportion dropped again in 1976, dropping from 53 percent in 1975 to 39 percent in 1976 and continuing a downward trend noted in 1974-75. This decrease in the proportion of M.A. students in economics offered aid, which is related undoubtedly to reduced university and departmental budgets, occurred in those departments which also offer Ph.D.s, both those in the Chairman's Group and the

[4]In 1976-77 for the fifth year, data related to supply of economists and academic demand for them are available from a survey of academic departments of economics. The data from the 1976-77 Universal Academic Questionnaires have been collected under the direction of C. Elton Hinshaw of the Association, and the data classified by sex are analyzed here. The questions asked in the 1976-77 survey are for the most part comparable to the data published in the Committee report in the May 1975 *Proceedings*. The number of departments which had reported in time for this analysis is 331 this year, but was 375 last year. Not all of the departments who reported last year reported again this year. Thus, comparisons of absolute numbers must be made with care. Percentages are more comparable, although, of course, they are subject to sampling error. Tabulations by sex from the 1977-78 survey are not available from the Association office in time to be included in this report.

[5]The 1974-75 comparison data quoted from the 1975-76 Universal Academic Questionnaires are from the Committee report, May 1976 *Proceedings*, pp. 512-20.

TABLE 5—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, 1975–76

Degrees Granted in 1975–76	All Departments	Highest Degree Offered: Ph.D. Chairman's Group	Ph.D. Other	M.A.	B.A.
Number of departments reporting	331	44	45	48	194
Ph.D., number	705	492	213	—	—
Percent women	10.4	10.4	10.3	—	—
M.A., number	1346	664	452	230	—
Percent women	13.4	12.0	15.3	13.5	—
B.A., number	9521	3921	1336	823	3441
Percent women	18.2	14.9	16.6	16.3	23.1
Other degrees from economics departments, number	70	34	36	0	3
Percent women	20.0	11.8	27.8	0	100.0

Source: Departments in United States and Canada reporting on 1976–77 Universal Academic Questionnaire.

other Ph.D. departments. Departments for which the M.A. degree is the highest degree offered in economics slightly increased the proportion of graduate students receiving aid (42 percent in the fall of 1975, 47 percent in the fall of 1976). Given this pattern of financial aid, the question is how women graduate students fared.

The proportion of women Ph.D. candidates receiving some financial aid

TABLE 6—NUMBER OF FULL-TIME "ON CAMPUS" GRADUATE STUDENTS REGISTERED FALL 1976, AND TYPE OF FINANCIAL AID, BY TYPE OF DEPARTMENT AND BY SEX

Type of Department, Degree Sought, and Sex	Total	Receiving Financial Aid: Tuition Only	Stipend Only	Tuition and Stipend	No Aid
All Departments					
Ph.D. students, number	2389	167	423	1212	587
Female as percent of total	14.3	19.2	9.2	15.6	13.8
M.A. students, number	1080	41	89	288	662
Female as percent of total	17.3	26.8	11.2	17.7	17.4
Chairman's Group					
Ph.D. students, number	1951	150	289	1026	486
Female as percent of total	14.5	16.7	10.7	15.6	13.8
M.A. students, number	570	19	36	134	381
Female as percent of total	15.6	21.1	8.3	16.4	15.7
Ph.D., other departments					
Ph.D. students, number	438	17	134	186	101
Female as percent of total	13.2	41.2	6.0	15.6	13.9
M.A. students, number	366	18	46	97	205
Female as percent of total	20.8	33.3	13.0	20.6	21.5
M.A. departments					
M.A. students, number	144	4	7	57	76
Female as percent of total	15.3	[a]	[a]	15.8	14.5

Source: See Table 5.
[a]Percentage not shown when fewer than 10 in cell.

continued in the fall of 1976, as in 1975, to be similar to or better than their proportionate representation among graduate students, except that the proportion of women Ph.D. candidates receiving stipend grants but not tuition in 1976 decreased, falling well below women's proportionate representation among graduate students. In contrast to the general favorable picture for tuition or tuition/stipend aid for women Ph.D. candidates, the proportion of women M.A. candidates receiving financial aid dropped sharply in the fall of 1976 compared with the previous fall. The proportion of women M.A. candidates receiving financial aid in the cartel departments dropped in 1976 compared with the previous year. However, the overall proportion of women M.A. candidates receiving financial aid in the Cartel departments was comparable to their proportionate representation among M.A. candidates, so that although the type of financial aid shifted, the overall cut in numbers of M.A. students offered financial aid was borne proportionately among the men and women studies who remained.

In general, based on tabulations of the approximately 1,600 women economists who have up-dated their current employment on the CSWEP roster of about 1,900 women economists, the type of employment in 1977–78 is shown in Table 7.

TABLE 7
(Shown in Percent)

	All Women Economists	With Ph.D.	With M.A. or B.A. as as Highest Degree
Total	100	100	100
Educational institution	59	77	47
Federal Government	5	4	5
State and Local Government	6	3	8
Quasi-Public Sector[a]	6	6	6
Consulting	12	7	16
Banking or finance	5	2	8
Industry	3	1	4
Students	4	0	6

[a]Often research institutions.

The first job of women after receiving their Ph.D.s in 1975–76 is shown in Appendix Table 4. The tabulation categories differ from those shown above. We know that women economists on the CSWEP roster include few women economists employed outside the United States and underreport the women in banking or finance, industry, and government, particularly women whose highest degree is an M.A. or B.A. Nevertheless, comparison of the first jobs of women after receiving Ph.D.s suggests that relatively fewer went into academic positions and relatively more went into government than was true of women economists as a whole. Of those women receiving M.A.s in economics in 1975–76, far fewer took teaching jobs, more went into industry, more continued as students, and about a third were employed outside the United States. If those employed outside the United States, the students, the unemployed and not known are excluded, the important relative shift of women with new M.A. degrees into industry is revealed.

Women economists in 1976 entering the labor market with a new M.A. or new Ph.D. still are not as apt to go into industry as their male classmates, and are more apt to go into academia. Men with new M.A. degrees in economics are more apt than women economists to be employed in federal or state or local government. Other differences in employment in 1976 between women and men with new degrees were small.

Considering all women economists employed in academic departments of economics, women in 1976–77 comprised 6 percent of the full-time faculty tenure-track positions; 14 percent of the full-time, nontenure-track positions, and 14 percent of the part-time faculty (Appendix Table 5). These proportions are similar to those reported for 1975–76, except the proportion of women in full-time, nontenure-track positions increased.

Within the full-time, tenure-track positions, the proportion of full professors was 5 percent in 1976–77 compared with only 3 percent the previous year. The increase occurred in Ph.D. department that are not in the Cartel and in departments in which the B.A. is the highest degree offered. There was a sharp drop in the proportion of women among instructors in all types of departments, with the sharpest drop in departments in the Chairman's Group. In these departments there was a corresponding increase in women reported in other faculty ranks and other positions. The increase in women at the assistant professor level, noted for 1975–76, leveled off in 1976–77.

The number of new faculty hired in 1976–77 exceeded the number of faculty released at the end of 1975–76 by 148 full-time positions and 43 part-time positions (Appendix Table 6). This represents a net increase of about 4 percent of the 4,070 full-time positions reported by the 331 departments participating in the 1976–77 survey, and 8 percent of the part-time faculty positions. The small net loss in professors and associate professors continues a pattern observed the previous year. The net increase in assistant professors hired was considerably larger than the previous year in spite of the fact that the number of departments participating was lower.

At each professorial rank, women tended to hold their own in these changes and even increased by 1 the number of full professors and associate professors at the same time there were more male retirements than new hires at those levels.

There was little difference between the prior type of economic employment of female and male economists hired in 1976–77 (Appendix Table 7). In the departments in the Chairman's Group, women economists were not hired from industry, banks or financial institutions, or the federal government, as were 7 percent of the men. In these departments, women were less apt than men to be hired from other university faculties. More than 60 percent of the women newly hired in these departments came straight from graduate school, as was also true for the male new hires. Of the female new hires, 15 percent in the Chairman's Group had previously been unemployed.

Women faculty released for 1976–77 were more apt than male economists to go to other faculty positions, and less likely than the men to go to business and industry, banking or financial institutions. This is a different pattern than reported in 1975–76.

In 1976–77 as in the two previous years, the persons reporting for the economics departments were asked to rank women full-time faculty by whether their salaries were above or below the departmental median for the particular rank and whether their length of service in that rank was above or below the median time at that rank for departmental faculty. Such estimates ignore how much the women's salary is above or below the median. From other evidence we know that with increases in experience, women's salaries tend to lag behind men's. For all departments, only 12 percent of the women had salaries more than $250 below the medians for their ranks (Table 8). When time in rank is considered, half of the women with salaries more than $250 below the median had time in rank at or above the median length of experience for that rank in the department. It must be remembered that two-thirds of the women faculty members in economics covered in the 1976–77 survey reported here are at the assistant professor or lower ranks. In general, entrance level faculty positions in universities have little or no difference between men and women in salary.

Women received 7 percent of the promotions for 1976–77 or 19 of the 256 (Table 9). Women comprised 8 percent of the total faculty. Of the 19 promotions for women, 4 were to full professor, 10 were to associate professor, and 5 were to assistant professor. None of the promotions of women to full professor included awarding of tenure. This may well be because the women already had tenure as associate professors. Eight women were awarded

TABLE 8—RELATIVE SALARIES FOR RANK AND TIME IN RANK OF FEMALE FULL-TIME ECONOMISTS, 1976–77, BY TYPE OF DEPARTMENT

Highest Degree Offered by Department and Relative Salary for Rank	All Women		Time in Rank[a]			
	Number	Percent	Total	Above Median	At Median	Below Median
All Departments	770	100.0				
Salary above median	383	49.7	100.0	38.4	35.2	26.4
Within $250 of median	298	38.7	100.0	34.2	52.0	13.7
Salary below median	89	11.6	100.0	41.6	10.1	48.3
Ph.D., Chairman's Group	53	100.0				
Salary above median	20	37.7	100.0	75.0	25.0	0
Within $250 of median	14	26.4	100.0	0	85.7	14.3
Salary below median	19	35.9	100.0	21.1	36.8	42.1
Ph.D., other	546	100.0				
Salary above median	315	57.7	100.0	35.2	33.3	31.4
Within $250 of median	214	39.2	100.0	46.7	52.8	0
Salary below median	17	3.1	100.0	17.6	5.9	76.5
M.A.	46	100.0				
Salary above median	18	39.2	100.0	50.0	44.4	5.5
Within $250 of median	14	30.4	100.0	14.3	57.1	28.6
Salary below median	14	30.4	100.0	28.6	7.1	64.3
B.A.	125	100.0				
Salary above median	30	24.0	100.0	40.0	56.7	3.3
Within $250 of median	56	44.8	100.0	0	39.3	60.7
Salary below median	39	31.2	100.0	66.7	0	33.3

Source: See Table 5.
[a]Shown in percent.

TABLE 9—PROMOTIONS AND TENURE DECISIONS FROM 1975–76 TO 1976–77 BY TYPE OF DEPARTMENT AND SEX

Highest Degree Offered by Department and Rank	Promotions to Rank		Given Tenure at Rank	
	Total Number	Female as Percent of Total	Total Number	Female as Percent of Total
All Departments				
Professor	86	4.7	11	0
Associate Professor	140	7.1	90	8.9
Assistant Professor	30	16.7	30	13.3
Ph.D., Chairman's Group				
Professor	27	3.7	2	0
Associate Professor	27	0	10	10.0
Assistant Professor	3	33.3	0	0
Ph.D., other				
Professor	23	0	3	0
Associate Professor	37	5.4	29	6.9
Assistant Professor	6	0	7	28.6
M.A.				
Professor	19	5.3	3	0
Associate Professor	35	5.7	22	9.1
Assistant Professor	6	16.7	9	11.1
B.A.				
Professor	17	11.8	3	0
Associate Professor	41	14.6	29	10.3
Assistant Professor	15	20.0	14	7.1

Source: See Table 5.

tenure at the associate professor level (compared with 10 promotions to this rank). Only 1 promotion and 1 tenure award to women in the professor and associate professor level were among departments in the Chairman's Group.

Critical to achieving improvements in the opportunities opened to women economists are actions by men of good will and sensitivity that will change traditionally narrow views of women's role potential and help open opportunities so women can have better educational and employment opportunities. Men and women economists must work together on this. Many of the improvements needed to combat role prejudice and sex discrimination in universities involve greater investment in on-the-job training opportunities for women and opening the informal network to women colleagues.

BARBARA B. REAGAN, *Chair*

APPENDIX TABLE 1—PROGRAM CHAIRS FOR ANNUAL AEA PROGRAM, BY SEX, 1969–77[a]

Year and Sponsor	Number of Sessions	Number of Chairs	Percent Female Chairs
1969			
AEA	24	24	8.3
Joint AEA	10	10	10.0
Graduate	1	1	0
Total	35	35	8.6
1970			
AEA	12	12	0
Joint AEA	18	16	0
Graduate	1	1	0
Total	31	29	0
1971			
AEA	21	22	13.6
Joint AEA	9	9	0
Graduate	1	1	0
Total	31	32	9.4
1972			
AEA	20	20	15.0
Joint AEA	13	13	15.4
Graduate	1	1	0
Total	34	34	14.7
1973			
AEA	15	15	6.7
Joint AEA	26	26	3.8
Graduate	1	1	0
Total	42	42	4.8
1974			
AEA	33	33	12.1
Joint AEA	11	11	9.1
Graduate	1	1	0
Total	45	45	11.1
1975			
AEA	31	30	10.0
Joint AEA	39	39	0
Graduate	1	1	0
Total	71	70	4.3
1976			
AEA	28	28	17.9
Joint AEA	21	21	4.8
Graduate	1	1	[b]
Total	50	50	14.0
1977			
AEA	46	46	6.5
Joint AEA	33	33	6.1
Graduate	1	1	0
Total	80	80	6.2

[a]Excludes presidential addresses and special lectures.
[b]Percentage not shown when fewer than 10 in cell.

Appendix Table 2—Author of Papers at Annual AEA Program, by Sex, 1969–77[a]

Year and Sponsor	Number of Papers	Number of Persons Writing Papers[b]	Percentage of Females[b] by Paper Given
1969			
AEA	72	85	3.5
Joint AEA	25	28	3.6
Graduate	4	4	0
Total	101	117	3.4
1970			
AEA	34	37	2.7
Joint AEA	59	78	10.3
Graduate	3	3	0
Total	96	118	7.6
1971			
AEA	45	49	8.2
Joint AEA	27	37	5.4
Graduate	4	4	[c]
Total	76	90	7.8
1972			
AEA	44	55	5.5
Joint AEA	34	44	4.5
Graduate	3	3	[c]
Total	81	102	5.9
1973			
AEA	24	30	10.0
Joint AEA	75	103	4.9
Graduate	4	5	0
Total	103	138	5.8
1974			
AEA	16	22	31.8
Joint AEA	125	153	9.8
Graduate	4	4	0
Total	145	179	12.3
1975			
AEA	97	109	13.8
Joint AEA	120	174	4.6
Graduate	3	3	[c]
Total	220	286	8.4
1976			
AEA	91	117	11.1
Joint AEA	64	73	12.3
Graduate	3	3	[c]
Total	158	193	11.9
1977			
AEA	146	200	10.0
Joint AEA	95	117	4.3
Graduate	3	3	[c]
Total	244	320	8.1

[a]Excludes presidential addresses and special lectures.
[b]Includes multiple authors.
[c]Percentage not shown when fewer than 10 in cell.

Appendix Table 3—Discussants of Papers at Annual AEA Program, by Sex, 1969–77

Year and Sponsor	Number of Discussants	Percent Discussants Female
1969		
AEA	71	2.8
Joint AEA	26	0
Graduate	4	[a]
Total	101	3.0
1970		
AEA	28	0
Joint AEA	34	0
Graduate	3	0
Total	65	0
1971		
AEA	70	11.4
Joint AEA	24	12.5
Graduate	3	0
Total	97	11.3
1972		
AEA	66	19.7
Joint AEA	37	5.4
Graduate	3	0
Total	106	14.2
1973		
AEA	48	12.5
Joint AEA	59	8.5
Graduate	0	—
Total	107	10.3
1974		
AEA	32	15.6
Joint AEA	65	10.8
Graduate	0	—
Total	97	12.4
1975		
AEA	64	10.9
Joint AEA	92	9.8
Graduate	3	0
Total	159	10.1
1976		
AEA	52	5.8
Joint AEA	54	7.4
Graduate	2	[a]
Total	108	7.4
1977		
AEA	79	5.1
Joint AEA	70	10.0
Graduate	3	[a]
Total	152	8.6

[a]Percentage not shown when fewer than 10 in cell.

Appendix Table 4—1976 Employment of 1975–76 Graduates in Economics by Level of Degree, Sex, and Type of Department

Type of Department and Kind of Employment	Ph.D.[a]		M.A.	
	Male	Female	Male	Female
All Departments:				
Number	532	66	383	58
Percent	89.0	11.0	86.8	13.2
Percent employed as economist in U.S.:				
Educational institution	53.9	57.6	5.5	12.1
Business or industry	2.8	0	13.6	6.9
Federal government	10.0	10.6	4.7	1.7
State/local government	2.1	4.5	8.9	1.7
Banking or finance	2.3	4.5	3.9	3.4
Consulting/research institution	3.2	4.5	1.3	1.7
Percent not employed as economist:				
Seeking employment	1.5	1.5	3.1	6.9
Not in labor force	4.5	6.1	1.0	5.2
Percent in other activities:				
Postdoctoral program	0	1.5	0	0
Entered the Ph.D. program	0	0	22.5	17.2
Employed outside U.S.	12.4	6.1	24.5	32.8
International Agency	3.2	0	0	0
Not known	4.1	1.5	11.0	10.3
Chairman's Group:				
Number	413	54	181	33
Percent	88.4	11.6	84.6	15.4
Percent employed as economist in U.S.:				
Educational institution	53.3	57.4	5.5	6.1
Business or industry	1.7	0	7.7	9.1
Federal government	11.6	13.0	2.8	3.0
State/local government	1.9	1.9	6.6	0
Banking or finance	2.2	5.5	2.8	3.0
Consulting/research institution	3.6	3.7	1.1	0
Percent not employed as economist:				
Seeking employment	1.2	0	2.2	6.1
Not in labor force	5.8	7.4	1.1	6.1
Percent in other activities:				
Postdoctoral program	0	1.9	0	0
Entered Ph.D. program	0	0	27.1	24.2
Employed outside U.S.	13.3	7.4	33.7	30.3
International Agency	3.9	0	0	0
Not known	1.5	1.9	9.4	12.1

Note: Percentages may not add to 100.0 due to rounding.
Source: See Table 5.
[a]Includes graduate students who have not completed their dissertations, if they entered the labor market seeking full-time employment as economists.

APPENDIX TABLE 5—NUMBER OF FACULTY BY RANK AND TYPE OF DEPARTMENT, 1976–77, BY SEX

Type of Appointment, Rank, and Sex	All Departments	Highest Degree Offered: Ph.D. Chairman's Group	Ph.D. Other	M.A.	B.A.
Departments reporting	331	44	45	48	194
Full-time faculty, tenure-track:					
All ranks, male and female	3841	1117	1193	572	959
Professors	1405	555	450	182	218
Associate professors	976	208	356	166	246
Assistant professors	1135	290	258	181	406
Instructors	156	17	67	13	59
Other faculty ranks	50	22	17	2	9
Other	119	24	45	28	21
Female percent of total	6.5	4.6	6.2	5.2	8.3
Professors	5.2	1.4	9.1	3.3	8.3
Associate professors	4.4	2.9	3.4	6.0	6.1
Assistant professors	7.7	10.0	5.0	6.1	8.6
Instructors	10.3	11.8	3.0	15.4	16.9
Other faculty ranks	18.0	18.2	17.6	[a]	11.1
Other	5.0	8.0	6.7	0	4.8
Full-time faculty, nontenure-track:					
All ranks, male and female	229	27	73	53	76
Professors	7	0	5	2	0
Associate professors	12	0	5	3	4
Assistant professors	85	12	28	17	28
Instructors	62	11	4	20	27
Other faculty ranks	26	2	8	6	10
Other	37	2	23	5	7
Female, percent of total	14.4	14.8	9.6	24.5	11.8
Professors	0	0	0	0	0
Associate professors	0	0	0	0	0
Assistant professors	15.3	0	14.3	23.5	17.8
Instructors	19.4	18.2	25.0	30.0	11.1
Other faculty ranks	19.2	[a]	[a]	[a]	0
Other	8.2	0	4.3	[a]	[a]
Part-time faculty:					
All ranks, male and female[b]	551	104	158	96	193
Professors	46	15	12	8	11
Associate professors	29	4	11	4	10
Assistant professors	74	8	20	15	31
Instructors	218	35	64	33	86
Other faculty ranks	110	28	28	20	34
Other	74	14	23	16	21
Female, percent of total[c]	14.3	14.4	15.8	12.5	14.0
Professors	4.3	6.7	0	0	9.1
Associate professors	10.3	0	18.2	0	10.0
Assistant professors	13.5	[a]	20.0	6.7	6.4
Instructors	13.3	8.6	12.5	18.2	14.0
Other faculty ranks	21.8	17.9	21.4	25.0	23.5
Other	14.9	21.4	21.7	0	14.3

Note: Percentages may not add to 100.0 due to rounding.
Source: See Table 5.
[a]Percentage not shown when fewer than 10 in cell.
[b]In all departments, 14 of these positions are tenure-track, 7 as professors. All 7 of the professors are in departments that are in the Chairman's Group.
[c]In all departments, 5 of these positions held by women are tenure-track, 1 at each rank.

APPENDIX TABLE 6—NET CHANGE IN FACULTY POSITIONS FROM END OF 1975-76 TO 1976-77, BY SEX, ALL DEPARTMENTS AND CHAIRMAN'S GROUP

Item	All Ranks	Professors	Associate Professors	Assistant Professors	Instructors	Other Faculty
All Departments:						
Faculty released end of AY 1975-75:[a]						
Full time, number	259	43	38	121	48	9
Women as percent of total	6.6	0	5.3	9.1	6.2	[b]
Part time, number	117	1	5	14	50	47
Women as percent of total	21.4	0	[b]	7.1	24.0	23.4
New Hires, faculty, AY 1976-77:						
Full time, number	407	22	32	251	81	21
Women as percent of total	10.8	4.5	9.4	10.0	16.0	9.5
Part time, number	160	4	1	29	88	38
Women as percent of total	15.0	0	0	6.9	13.6	26.3
Net change, 1975-76 and 1976-77:						
Full time, number	+148	−21	−6	+130	+33	+12
Women, number	+ 27	+ 1	+1	+ 14	+10	+ 1
Part time, number	+ 43	+ 3	−4	+ 15	+38	− 9
Women, number	− 1	0	−1	+ 1	0	− 1
Chairman's Group:						
Faculty released end of AY 1975-76:[a]						
Full time, number	79	21	13	35	6	4
Women as percent of total	5.1	0	15.4	5.7	0	0
Part time, number	44	0	0	3	19	22
Women as percent of total	18.2	0	0	0	21.1	18.2
New Hires, Faculty, AY 1976-77:						
Full time, number	107	12	5	71	14	5
Women as percent of total	4.7	0	0	4.2	7.1	[b]
Part time, number	39	0	0	0	31	8
Women as percent of total	5.4	0	0	0	12.9	[b]
Net Change, 1975-76 and 1976-77:						
Full time, number	+ 28	− 9	−8	+ 36	+ 8	+ 1
Women, number	+ 1	0	−2	+ 1	+ 1	+ 1
Part time, number	− 5	0	0	− 3	+12	−14
Women, number	− 2	0	0	0	0	− 2

Source: See Table 5.

[a]Resignation, retirement, and nonrenewal of contracts; AY denotes academic year.

[b]Percentage not shown when fewer than 10 in cell.

APPENDIX TABLE 7—PRIOR ACTIVITY OF NEW 1976-77 APPOINTMENTS AND PRESENT ACTIVITY OF "RELEASES" FOR 1975-76, BY TYPE OF DEPARTMENT AND SEX (Shown in Percent)

Highest Degree Offered by Department and Activity of Faculty	New Hires in 1976-77[a] by Prior Year Activity		Those Released for 1976-77 by Present Activity[a]	
	Male	Female	Male	Female
All Departments	100.0	100.0	100.0	100.0
U.S. business and industry	2.9	3.0	9.0	0
Fed./state government in U.S.	4.2	4.5	9.3	7.7
Outside U.S.	2.7	0	7.1	7.7
Faculty at another school	26.3	23.9	37.7	50.0
Bank or finance institution	1.0	0	4.1	0
Research institution	2.4	3.0	6.0	0
Graduate student	54.8	55.2	9.0	7.7
Postdoctoral program	1.2	1.5	1.1	3.8
Unemployed	0	4.5	1.9	0
Unknown	0	0	3.0	3.8
Other	4.2	4.5	11.9	19.2

Appendix Table 7—*(Continued)*

Highest Degree Offered by Department and Activity of Faculty	New Hires in 1976–77[a] by Prior Year Activity		Those Released for 1976–77 by Present Activity[a]	
	Male	Female	Male	Female
Chairman's Group	100.0	100.0	100.0	100.0
U.S. business and industry	1.0	0	6.4	0
Fed./state government in U.S.	4.6	0	8.5	0
Outside U.S.	1.5	0	8.5	a
Faculty at another school	22.5	15.4	41.5	a
Bank or finance institution	1.0	0	7.4	0
Research institution	3.1	7.7	5.3	0
Graduate student	63.5	61.5	9.6	0
Postdoctoral program	1.5	0	0	0
Unemployed	0	15.4	0	0
Unknown	0	0	1.1	0
Other	1.5	0	11.7	a
Ph.D., other	100.0	100.0	100.0	100.0
U.S., business and industry	8.0	0	8.7	0
Fed./state government in U.S.	3.4	7.1	11.3	a
Outside U.S.	4.5	0	8.8	a
Faculty at another school	23.9	14.3	30.0	a
Bank or finance institution	2.3	0	3.8	0
Research institution	2.3	0	10.0	0
Graduate student	45.5	71.4	5.0	0
Postdoctoral program	3.4	0	2.5	a
Unemployed	0	0	1.2	0
Unknown	0	0	1.3	0
Other	6.8	7.1	17.5	0
M.A.	100.0	100.0	100.0	100.0
U.S. business and industry	0	0	15.4	0
Fed./state government in U.S.	4.8	8.3	7.7	0
Outside U.S.	0	0	7.7	0
Faculty at another school	29.0	33.3	38.5	a
Bank or finance institution	0	0	0	0
Research institution	0	0	0	0
Graduate student	56.5	58.3	11.5	a
Postdoctoral program	0	0	3.8	0
Unemployed	1.6	0	3.8	0
Unknown	0	0	0	0
Other	8.1	0	11.5	a
B.A.	100.0	100.0	100.0	100.0
U.S. business and industry	3.1	7.1	10.3	0
Fed./state government in U.S.	3.9	3.6	8.8	10.0
Outside U.S.	3.9	0	2.9	0
Faculty at another school	30.5	28.6	41.2	50.0
Bank or finance institution	1.0	0	1.5	0
Research institution	3.1	3.6	4.4	0
Graduate student	51.6	42.8	11.8	10.0
Postdoctoral program	0	3.6	0	0
Unemployed	0	3.6	4.4	0
Unknown	0	0	8.8	10.0
Other	3.1	7.1	5.9	20.0

Note: Percentages may not add to 100.0 due to rounding.
Source: See Table 5.
[a]Percentage not shown when fewer than 10 in cell.

Report of the Economics Institute's Policy and Advisory Board

A strong and growing demand continues for the Institute's main services to the profession, namely the provision of an intensive transitional training program for entering graduate students from abroad. Enrollment in the 1977 Summer Session reached a record level of 249. These came from 49 different countries and continued at 53 universities.

The Institute has expanded its program to provide intensive English training for specialists in economics and related fields throughout the year. As a result, it is better able to help departments in the admissions process by accommodating students with lower beginning levels of proficiency in English for appropriate longer periods of preparatory training. The extended program also services participants planning spring semester admissions to graduate schools. An expanded component of supplementary training opportunities in mathematics, statistics, and economic theory is also being developed in the fall and spring programs to complement the English program and to provide for progressive diversification of training and a broader base of preparation for the Institute's Comprehensive Summer Session. The curriculum has also been expanded to include preparatory course work in accounting and a number of seminars and supplementary short courses on such topics as economic planning and development, American business systems, agricultural systems and development, management of development projects, financial systems, and computer fundamentals.

In addition to its basic training activities, the Institute has expanded its capacity to assist students and sponsoring agencies in university admissions procedures and to provide universities with professional evaluations of applicants. Complementary to these services, it has recently published an expanded edition of the *Guide to Graduate Study in Economics and Agricultural Economics*. Copies of this more recent edition (1977) are being distributed through the Irwin Publishing Co. at a cost of $11.50.

As a result of expanding enrollments and support by an increasing number of sponsoring agencies, both domestic and abroad, the Institute has been able to spread a five-year support grant provided to it by the Ford Foundation in 1968 over a ten-year period. It is, however, now seeking additional funding at this time to replenish its now seriously depleted general support funds, both to help perpetuate its continued operation and to expand its capacity to accommodate a significant proportion of the applicants who do not have alternative funding sources for their entire costs.

The Institute continues to recruit its faculty on a national basis and seeks to expand close working relationships with universities and sponsoring agencies in support of its efforts to improve the efficiency and effectiveness of foreign student selection and training. At a time when foreign students constitute between 30 and 40 percent of enrollment of most graduate programs in economics and agricultural economics throughout the United States, this is an extremely important endeavor. A growing list of distinguished alumni of the Institute attests to the positive role it is playing.

The Policy and Advisory Board met in Boulder during the summer. Arnold Harberger and Richard H. Holton completed three-year terms and have been suceeded by Carlos Diaz-Alejandro and Raymond Vernon.

EDWIN S. MILLS, *Chair*

Report of the Representative to the National Archives Advisory Council

On May 25, 1977, the Administrator of the General Services Administration, Joel W. Solomon, issued an order reconstituting the National Archives Advisory Council. The principal changes introduced by the order are as follows: In the future, the Council will advise the Archivist of the United States, rather than the General Services Administrator, as formerly. The Administrator will continue to appoint members of the Council, as before. However, in the future each organization represented on the Council will be expected to provide the Administrator with a slate of three to five nominees, from which the Administrator will choose a representative. The slate "should be balanced to ensure representation by women and minorities." Terms on the Council will continue to be three years in duration (beginning the first of the year) and future appointees will be permitted to serve no more than two terms.

My present term on the Council will end on December 31 of next year. Therefore within the next year the American Economic Association will be obliged to forward a slate of nominees to Mr. Solomon. Since the activities of the Council will be of interest chiefly (but not exclusively) to economic historians, I urge that the Board of Trustees of the Economic History Association be consulted in this matter. The appropriate person to bring the subject before the Board of Trustees is the President of the Association, Lance Davis, of the California Institute of Technology.

The Council met twice in 1977, on May 12–14 and December 1–3, at the National Archives in Washington. The group was briefed on and/or provided advice with respect to the following topics: The Nixon papers and tapes, plans for the Gerald R. Ford Library (Ann Arbor) and Museum (Grand Rapids), plans for the new Archives building in Washington, the Kissinger papers and tapes, the Public Documents Commission, the Privacy Act, the Freedom of Information Act, the State Department records system, machine readable records and the Archives problems of paper and film records preservation, the National Records Centers, NARS training activities, the new Executive Order on classification and declassification of National Security Documents, and the 1900 manuscript census records. While I will be happy to supply further information on any of these topics, I do not think the Executive Committee requires further briefing on any of them—with one exception—nor do I think that any action on the part of the Committee is currently required. The one exception is the 1900 manuscript census.

As I have previously reported to the Executive Committee, the manuscript records of the 1900 population census have been opened, under certain restrictions established by the Archivist of the United States at the insistence of the Census Bureau. On November 1 of this year, the Archivist advertised in the Federal Register his intention of freeing access to the records from the previously established restrictions. Since no objections have been forthcoming, the Archivist's new order will be in force in future. This means that early next year the Archives will be in a position to sell microfilms of the 1900 population manuscript records. These records contain demographic information (at the level of the household), together with some economic and social information (occupation, education, literacy, home-ownership). As soon as the details of the sale are established I will notify the editor of the *American Economic Review*, in the hope that he will publish a note advising the members of the Association of this new data source.

R. E. GALLMAN, *Representative*

Report of the Committee on U.S.-Soviet Exchanges

A significant development this year was the second meeting of the Commission on the Social Sciences and Humanities of the Academy of Sciences of the *USSR* and the American Council of Learned Societies, held in Moscow on June 6–10, 1977. The Protocol signed at that time lists the American Economic Association exchange program as one of the agreed upon activities (Sub-Commission on Economics, Item III: "Organization of joint U.S.–U.S.S.R. symposium on major problems of the U.S.A. and the U.S.S.R. in economics"). This makes the program eligible, in principle, for official funding and for logistical aid from IREX (International Research and Exchanges Board). In actuality, however, there is no assured source of funds, and each of the three symposia held thus far has been financed on a hand-to-mouth basis.

Our last report described a proposal for the Association to take over from the National Bureau of Economic Research the administration of the National Bureau's program on application of computers to economic management. This plan has not materialized. The National Bureau will probably be withdrawing from this program in the near future; but it appears that such activities as remain will be administered directly by the computer sciences division of the National Science Foundation through individual grants to *U.S.* scholars.

Under the symposium program, eight Soviet economists were invited to visit the United States for the period November 27–December 11, 1977, of whom only six arrived. The third symposium meeting, "Aspects of Labor Economics" was held at Wingspread, the Johnson Foundation Conference Center in Racine, Wisconsin, on November 28–30.

The participants felt that the meeting was interesting and productive. The *U.S.* papers, and several of the Soviet papers, were of high quality. Especially interesting was the round table discussion of the papers, which was conducted in a frank and friendly spirit, and which brought out many sidelights on the operation of the two economies.

After the symposium, the Soviet group spent several days each in San Francisco, Washington, and New York. They visited two universities, two major banks, three industrial plants; and in Washington they met with the Secretary of Labor, the Secretary of the Treasury, members of the Council of Economic Advisers, and staff members of the Joint Economic Committee. The itinerary was laid out with a view to eliciting comparable treatment for future *U.S.* groups visiting the *USSR*.

Funding for the Soviet visit was provided by IREX, which also handled hotel and plane reservations and other logistical matters. We are grateful to President Allen Kassof and other members of the IREX staff not only for financial assistance but also for their efficient administrative arrangements, which reduced the burden falling on our Committee and on the Association's headquarters.

Toward the close of the Soviet visit, members of our Committee met with President Klein in New York to discuss possible future activities. We then met with Academician Khachaturov, the Soviet coordinator, and two of his colleagues for joint discussion. At this meeting it was agreed that eight American economists would be invited to the *USSR* for two weeks in June 1978, the visit to include a symposium in Moscow on problems of industrial management. Our Committee has given some thought to possible paper topics and *U.S.* participants, but we would welcome advice from the Executive Committee on these matters.

We believe that the program has proven fruitful enough to be continued through another two-year round of exchanges. The June 1978 visit will provide an opportunity to discuss possible subjects for future symposia. Topics which have been suggested tentatively from our side or from the Soviet side include: economics of natural

resources; economics of the energy industries; economics of agriculture; economic analysis of environmental protection; price formation and functions of prices; distribution of income. Again, we would welcome advice from the Executive Committee on the priority which should be given to these or other topics.

Concerning other possible future activities, we have one suggestion. Through the exchange visits to date, we have developed good personal relations with directors of several of the research institutes in Moscow—Institute of Economics, Institute of World Economics and International Relations, Economics-Mathematics Institute, Institute of the U.S.A. and Canada, Institute of Labor—and also with the institute directors in other places such as Kiev, Leningrad, and Alma Ata. These contacts could perhaps be used to promote research cooperation of a more serious sort than can be accomplished in brief symposia. Specifically, there is a possibility of developing programs of parallel research, in which one or more Soviet economists and one or more American economists would work simultaneously at their home institutions on similar problems, meeting occasionally for interchange of ideas and findings. The Association, with its limited staff, would not wish or need to get deeply involved in administering such activities. But it might play a useful role in authenticating and sponsoring joint programs, and in identifying institutions and individuals willing to participate in them.

This possibility has not been discussed with our Soviet colleagues, who might or might not be interested. We could float some trial balloons in Moscow next June, however, if this seems expedient.

Lack of assured funding for our exchange activities is a continuing problem. We propose to continue exploring possible government and private sources, in collaboration with IREX. Advice and assistance on this front will be especially welcome.

LLOYD G. REYNOLDS, *Chair*

Report of the Representative to the National Bureau of Economic Research

In 1977 there was a transition of the National Bureau leadership. Martin Feldstein became president in April, succeeding John Meyer who had served as president since 1967. Several new program directors were appointed as part of the general restructuring of the research program that began this year. The new programs and their directors are: Economic Fluctuations (Robert Hall); Financial Markets and Monetary Economics (Benjamin Friedman); Business Taxation and Finance (David Bradford); Labor Markets (Richard Freeman); Social Insurance (Michael Boskin); and Growth of the American Economy (Robert Fogel). Charles McLure, Jr. assumed the new position of Executive Director for Research, located in the office of the president in Cambridge, Massachusetts.

The Computer Research Center under the direction of Edwin Kuh will be transferred in 1978 from the National Bureau to the Massachusetts Institute of Technology.

Research programs were continued in the general areas of Human Behavior and Social Institutions (economics of health, population, family, law and legal institutions, income distribution and labor markets) under the direction of Victor Fuchs; International Studies, and Financial Institutions and Industrial Organizations, under Robert E. Lipsey; and Business Cycles, Prices and Productivity under Geoffrey H. Moore.

Board of Directors: James J. O'Leary was elected Chairman of the Board, succeeding J. Wilson Newman. Eli Shapiro was elected Vice Chairman, succeeding Moses Abramovitz.

At the annual meeting of the Board in September, Moses Abramovitz, Richard N. Rosett, and Bert Seidman were elected as new Directors at Large. George Leland Bach succeeded Abramovitz as Director from Stanford University. Rudolph A. Oswald was elected as the representative Director from the AFL-CIO and Philip J. Sandmaier, Jr. as the Director from the American Institute of Certified Public Accountants. Emilio G. Collado and Thomas D. Flynn were elected Directors Emeriti. In December James L. Pierce replaced Daniel L. McFadden as Director from the University of California-Berkeley.

Conferences and Workshops: The Conference on Research in Income and Wealth met on December 8–9, 1977 in Williamsburg, Virginia and dealt with Modeling the Distribution and Intergenerational Transmission of Wealth.

The Universities-National Bureau Committee for Economic Research has planned a conference in late spring 1978 on Low Income Labor Markets with Sherwin Rosen as chairman, and a conference to be held later in the year on Economics of Information and Uncertainty with John McCall as chairman. A miniconference on Corporate Financial Policy under the direction of David Bradford was held in Cambridge on December 14, 1977.

The Latin American Computer Workshops program included a jointly sponsored NBER-ECIEL conference (Programa de Estudios Conjuntos sobre Integracion Economica Latinoamericana) on September 19–21, 1977 in Washington, D.C. to deal with Education and Economic Development in Latin America. Another conference on Commodity Markets, Models and Policies in Latin America is scheduled to be held in 1978 in Lima, Peru.

Two conferences with National Bureau sponsorship were held on the general topic of Alternatives for Growth: one cosponsored by Clemson University was addressed to the Engineering and Economics of Natural Resource Development, and met at Clemson, S.C., on April 4–5; the other, on Urban Development, was cosponsored by the Joint Center for Urban Studies of

MIT and Harvard, and met in Cambridge, Mass., June 9–10, 1977.

A Conference on Public Regulation, sponsored by the National Bureau under a grant from the National Science Foundation/Research Applied to National Needs, was held on December 15–17, 1977 in Arlington, Virginia.

A National Bureau-sponsored workshop on stochastic control theory was organized by David Kendrick and Edison Tse, and held at Yale University on May 25–27, 1977.

The National Bureau and the Social Security Administration are jointly sponsoring a workshop on policy analysis with social security research files. It is being organized by Frederick Scheuren of the Social Security Administration and Richard C. Taeuber of the Commission on Federal Paperwork, and is scheduled to be held in Williamsburg, Virginia, March 15–17, 1978.

James Heckman is chairman of a planned conference on uses of longitudinal data which is scheduled to be held in June 1978.

Research Fellows: For 1976–77 National Bureau Faculty Research Fellows were Daniel A. Graham of Duke University, Cheng Hsaio of the University of California-Berkeley, and J. Huston McCulloch of Boston College. Narongchai Akrasanee of Thammasat University, Bangkok was a Foreign Research Fellow in 1976–77. Research Fellows for 1977–78 are Michael D. Hurd of Stanford University, Mieko Nishimizu of Princeton University, and Daniel A. Seiver of the University of Alaska.

Publications: Six National Bureau books were published in 1977: *Foreign Trade Regimes and Economic Development: Colombia*, Carlos E. Diaz-Alejandro; *Foreign Trade Regimes and Economic Development: Chile*, J. R. Behrman; *Substituting A Value-Added Tax for the Corporate Income Tax*, S. P. Dresch, An-loh Lin, and D. K. Stout; *Education As An Industry*, J. N. Froomkin, D. T. Jamison, and R. Radner, eds.; *Analysis of Inflation: 1965–74*, Joel Popkin, ed.; *Residential Location and Urban Housing Markets*, C. K. Ingram, ed.

The National Bureau also published the quarterly journals *Explorations in Economic Research* and *Annals of Economic and Social Measurement*. It is expected that publication of these journals by the National Bureau will be discontinued in 1978.

During 1977 four issues were published of a new informational bulletin, *NBER Reporter*, which provides brief accounts of current activities and research reports. Anyone wishing to receive a free subscription should write to the National Bureau of Economic Research, 1050 Massachusetts Ave., Cambridge, Massachusetts, 02138.

Further information on research projects and related activities during the past year is contained in the *NBER Annual Report*, September, 1977, available on request.

CARL F. CHRIST, *Representative*

Report on Joint *Ad Hoc* Committee on Government Statistics

The Joint *Ad Hoc* Committee on Government Statistics (JAHCOGS) was established in 1975 by five professional associations in response to a call from the American Statistical Association. It was convened in the belief that the federal statistical system failed in several important ways to measure up to standards that professionals found both feasible and necessary. In 1976 the Committee issued a preliminary report which described several problem areas and presented recommendations for change. In the spring of 1977 the Committee expanded its membership to nine associations, including the American Economic Association.

In its final report, which will be available by the end of March, the enlarged Committee presents more specific commentary and recommendations on the problems it has identified, together with its recommendation that the associations establish instrumentalities through which their members can more readily monitor developments in federal statistics.

In order to help accomplish this objective, JAHCOGS recommends that:

1) There be established a Committee of Professional Associations on Federal Statistics.

2) There be established an Office of Professional Associations on Federal Statistics under the policy guidance of the Committee. This office is to be administratively located within the Federal Statistics Users' Conference (FSUC) and to be administered by the Executive Director of FSUC.

3) Each association now participating in JAHCOGS, and others that may be added, contribute $2,000 per year to the costs of operation of the Office of Professional Associations on Federal Statistics for an initial three-year trial period.

As the American Economic Association's representatives to JAHCOGS we endorse these recommendations and request the Executive Committee to:

1) Join the Committee of Professional Associations on Federal Statistics and appoint a member of the Association to serve as its representative on the Committee.

2) Make the necessary budget allocation to support the work of the Office of Professional Associations on Federal Statistics during the trial period.

Member Organizations of JAHCOGS

- American Economic Association
- American Political Science Association
- American Public Health Association
- American Sociological Association
- American Statistical Association
- Federal Statistics Users' Conference
- National Association of Business Economists
- Population Association of America
- Society of Actuaries

EDWARD F. DENISON
GARY FROMM
Representatives